NAVAL HEALTH RESEARCH CENTER

DEVELOPMENT OF A SAMPLING STRATEGY FOR DISEASE AND NON-BATTLE INJURY (DNBI) DATA RATES

I. T. Show

M. R. White

Technical Document 95-2B

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92186 – 5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

DEVELOPMENT OF A SAMPLING STRATEGY FOR DISEASE AND NON-BATTLE INJURY (DNBI) DATA RATES

Prepared for:

NAVAL HEALTH RESEARCH CENTER
P.O. Box 85122
San Diego, CA 92138

Prepared By:

Ivan T. Show, Ph.D
SOUTHWEST RESEARCH ASSOCIATES, INC.
2006 Palomar Airport Road, Suite 207
Carlsbad, CA 92008

and

Martin R. White, M.P.H.
NAVAL HEALTH RESEARCH CENTER
San Diego, CA 92186-5122

Report No. 95-2B Supported by the Navy Medical Research and Development Command under work unit M0095.005-6103, Department of the Navy. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Navy, Department of Defense, or United States Government.

Summary

Problem: Military health-care planners need to have the capability to easily detect, sample, and analyze any ICD9-CM illness or injury based upon a number of ancillary variables including age, race, sex, service branch, ship type, pay grade, and occupation if they are to provide the best health-care possible.

Objective: To provide military health-care planners with an integrated system of computer programs that will significantly improve the ability to access various medical files, allowing for the continued monitoring of the health and medical needs of U.S. military personnel.

Approach: A system of computer programs is being developed using the latest in object-oriented technology designed to run on any IBM-compatible PC under the Windows O/S. The system (EPISYS) consists of a number of user modules, including EPILIMIT, EPIBASE, EPISAAM, EPIMIPS, and a Utilities module, designed to work as an integrated system. These modules each provide a useful function, and additional modules can be added as required.

Results: The system, when complete, will give researchers and medical planners the capability to easily detect, sample, and analyze any ICD9-CM illness based upon a number of ancillary variables including age, race, sex, service branch, ship type, pay grade, and occupation. These programs will significantly improve the ability to access various medical files, and will provide investigators with an integrated system of computer programs for health monitoring and medical projection needs.

Conclusions: Providing medical researchers and military health-care planners with a system of computer programs to rapidly access medical information greatly assists them with making research, clinical, and management decisions. The work to integrate EPISYS, and other medical systems will be studied in an effort to expand the utility of the system.

TABLE OF CONTENTS

1.0 INTRODUCTION

The goal of this project was to develop a statistical sampling scenario suitable for determining Disease and Non-Battle Injury (DNBI) rates. The target population was the United States Navy (USN) and Marine Corps (USMC). The purpose of the project was to develop methods by which DNBI rates could be analyzed, projected, and compared among USN and USMC units. Specifically, the methods developed were to be capable of detecting a two-fold increase or decrease between DNBI rates for different units during the same time period or for the same units between different time periods.

The Naval Health Research Center (NHRC) must deal with the realities of projecting DNBI rates that are scientifically and statistically meaningful as well as being useful at USN staff and command levels. The primary reason for reporting DNBI rates is to project medical-care requirements for peacetime operations, for wartime redeployment, and for determining operational readiness. For this reason, it was recognized that relatively simple and straightforward representations of DNBI rates were required. Therefore, major problems were how to reduce the sheer number of International Classification of Diseases, Revision 9, Clinical Modification (ICD9-CM) reporting categories and how to manage the complexity of the USN and USMC populations.

Several criteria for sampling methods were developed according to overall NHRC requirements. Based on these requirements, it was decided that the sampling scenario would possess the following:

- the ability to perform routine monitoring in real time;
- statistically meaningful predictive capability;
- enhanced usefulness of existing data sources;
- improved framework for future experimental designs and sampling.

During the course of the project, a statistical sampling scenario possessing the desired criteria was developed. The statistical developments were incorporated into a comprehensive, user-friendly computer program. The program automates the task of analyzing historical data and determining current sample size requirements. In addition, it provides both screen and hard-copy output for all technical and intermediate results.

2.0 DESCRIPTION OF STATISTICAL SAMPLING PLAN

2.1 Overview

This section provides a summary overview of the statistical sampling plan. All variables are defined in section 2.2, methods in sections 2.3 through 2.8.

The statistical sampling plan is defined in a multivariate context, meaning that there is more than one response variable. In fact, the basic data set is a complex, six-dimensional matrix defined by two independent variables (ICD9-CM reporting categories and month of the year) and four ancillary variables (service branch, pay grade, platform, and location). Response variables are the number of DNBI incidence rates within each cell of the data matrix.

Because of the large number of variable combinations, the first step is to reduce the number of independent variables. This reduction is accomplished by a classification analysis on a matrix of DNBI incidence rates. The axes of the matrix are ICD9-CM categories and months of the year. The purpose of the analysis is to place ICD9-CM categories and months into internally consistent classes. The classes then become the basis of further analyses, thus reducing the number of the variables without any significant loss of information.

The second step, multiple discriminant analysis, accomplishes two things. First, it places the ancillary variables into groups based on their relationships with the previously defined independent variable classes. Second, it is used to define sampling strata based on the ancillary variable groups.

The third step is a time series analysis. The sample allocation method used requires a knowledge of the fundamental periods at which DNBI rates vary over time. Therefore, main time series method is calculation of power spectral densities for various partitions of the basic data set and subsequent extraction of the fundamental periods.

The final step is the calculation of sample sizes. First, sampling strata are defined based on the ancillary variable groups. Then, using a form of optimum allocation, sample sizes are calculated for each stratum. Sample sizes are calculated for the overall database or for individual ICD9-CM categories.

2.2 Definitions

Six variables define the structure of the basic data set. Two variables are termed independent variables: (1) ICD9-CM reporting categories, and (2) month of the year. The other four variables are termed ancillary variables: (1) service branch, (2) pay grade, (3) platform, and (4) location.

ICD9-CM reporting categories are defined and used in two different but related manners. For the main multivariate analyses there are 29 major categories. For univariate sample size allocations there are 130 detailed categories. In Table 1, the univariate categories are listed under the appropriate multivariate categories.

TABLE 1. ICD9-CM CATEGORIES
Viral Diseases
Measles
Rubella
Chickenpox
Herpes Zoster
Herpes Simplex
Smallpox
Acute Poliomyelitis
Aseptic Meningitis (Enterovirus)
Other Enterovirus Diseases of CNS
Cowpox
Other Viral Exanthemata
Yellow Fever
Dengue
Mosquito-borne Viral Encephalitis
Tick-borne Viral Encephalitis
Unspecified Arthropod-borne Encephalitis
Acute Encephalitis Epidemic
Viral Encephalitis
Arthropod-borne Hemorrhagic Fever
Epidemic Hemorrhagic Fever
Other Arthropod-borne Viral Diseases
Viral Hepatitis
Rabies
Mumps
Coxsackie Virus Disease
Infectious Mononucleosis
Other Viral Diseases of Conjunctiva
Other Viral Diseases
Bacterial Diseases
Cholera
Typhoid Fever
Paratyphoid Fever
Other Salmonella Infections
Bacillary Dysentery
Bacterial Food Poisoning

TABLE 1. (continued)

Bacterial Diseases (continued)
Diarrheal Diseases
Plague
Tularemia
Anthrax
Brucellosis
Glanders
Melioidosis
Ratbite Fever
Other Zoonotic Bacterial Diseases
Leprosy
Diphtheria
Whooping Cough
Streptococcal Sore Throat, Scarlet Fever
Erysipelas
Meningococcal Infection
Tetanus
Septicemia
Other Bacterial Diseases
Mycobacterial Diseases
Tuberculosis Pulmonary
Tuberculosis Other Respiratory
Tuberculosis Meninges CNS
Tuberculosis Intestine, Peritoneum, Mesentery
Tuberculosis Bones, Joints
Tuberculosis Genitourinary
Tuberculosis Other Organs
Tuberculosis Disseminated
Tuberculosis Late Effects
Tuberculosis Unspecified Site
Other Mycobacterial Diseases
Rickettsial Diseases
Tick-borne Rickettsiosis
Psittacosis
Other Rickettsiosis
Louse-borne Typhus
Chlamydial Diseases
Trachoma
Sexually Transmitted Diseases
Syphilis Congenital
Syphilis Early, Symptomatic

TABLE 1. (continued)

Sexually Transmitted Diseases (continued)
Syphilis Early, Latent
Syphilis Cardiovascular
Syphilis of CNS
Syphilis Other and Unspecified
Syphilis
Gonococcal Infections
Other Venereal Diseases
Trichomoniasis Urogenital
Fungus Diseases
Dermatophytosis
Dermatophytosis Other and Unspecified
Moniliasis
Actinomycosis
Coccidioidomycosis
Blastomycosis
Other Mycoses
Diseases Caused by Spirochetes
Relapsing Fever
Leptospirosis
Vincent's Angina
Yaws
Pinta
Other Spirochete Infections
Protozoal Diseases
Malaria
Leishmaniasis
American Trypanosomiasis
Other Trypanosomiasis
Toxoplasmosis
Amebiasis
Other Protozoal Intestinal Diseases
Ectoparasites
Pediculosis
Acariasis
Other Infestations
Other Arthropod Infestations
Diseases Caused by Worms
Schistosomiasis
Other Trematode Infections
Hydatidosis

TABLE 1. (continued)

Diseases Caused by Worms (continued)
Other Cestode Infections
Trichiniasis
Filarial Infection
Ancylostomiasis
Other Intestinal Helminthiasis
Other and Unspecified Helminthiasis
Intestinal Parasitism Unspecified
Neoplasms
Endocrine, Nutritional, and Metabolic Diseases
Diseases of Blood and Blood-Forming Organs
Mental Disorders
Nervous System and Sense Organs
Diseases of the Circulatory System
Diseases of the Respiratory System
Diseases of the Digestive System
Diseases of the Genitourinary System
Complications of Pregnancy, Childbirth, Puerperium
Diseases of Skin and Subcutaneous Tissue
Diseases of Musculoskeletal System and Connective Tissue
Congenital Anomalies
Perinatal Morbidity and Mortality
Symptoms and Ill-Defined Conditions
Accidents, Poisonings, and Violence
Supplementary Classification Special Conditions

The time variable is month-of-the-year; the realizations are simply January through December. The first ancillary variable is service branch; its realizations are U.S. Navy and U.S. Marine Corps. The second ancillary varialbe is pay grade and its realizations are enlisted and officer (including warrant).

The third ancillary variable is platform; there are 47 platform definitions. They are listed in Table 2. The fourth ancillary variable is location; its 14 realizations are listed in Table 3.

TABLE 2. PLATFORMS

Ashore Stations	
AE	Ammunition Ship
AF	Combat Stores Ship
AFS	Combat Stores Ship
AGF	Command Ship
AGS	Survey Ship
AGSS	Survey Ship, Small
AO	Oiler
AOE	Fast Combat Support Ship
AOR	Replenishment Oiler
AR	Repair Ship
ARL	Repair Ship
ARS	Repair Ship, Small
AS	Submarine Tender
ASR	Submarine Rescue Ship
ATF	Fleet Tug
BB	Battleship
CA	Cruiser, Gun
CG	Cruiser, Guided Missile
CGN	Cruiser, Guided Missile, Nuclear
CV	Aircraft Carrier
CVA	Aircraft Carrier, Assualt
CVN	Aircraft Carrier, Nuclear
CVS	Aircraft Carrier
DD	Destroyer
DDG	Destroyer, Guided Missile
FF	Frigate
FFG	Frigate, Guided Missile
LCC	Amphibious Assault Ship, Command
LHA	Amphibious Assault Ship, General Purpose
LHD	Amphibious Assault Ship, Multipurpose
LKA	Amphibious Cargo Ship
LPD	Amphibious Transport, Dock
LPH	Amphibious Assault Ship, Helicopter
LSD	Landing Ship, Dock
LST	Landing Ship, Tank
MCM	Mine Countermeasures Ship
MSO	Minesweeper
SS	Submarine, Attack, Diesel
SSBN	Submarine, Ballistic Missile, Nuclear
SSN	Submarine, Attack, Nuclear

TABLE 3. LOCATIONS

A1 - Atlantic, Northwest	P1 - Pacific, Northeast
A2 - Atlantic, South	P2 - Pacific, Central
A3 - Atlantic, Northeast	P3 - Pacific, West
A4 - Caribbean & Gulf of Mexico	P4 - Pacific, Southeast
A5 - Arctic Ocean	P5 - Bering Sea
A6 - Mediterranean Sea	P6 - Antarctic Ocean
A7 - Persian Gulf & Red Sea	P7 - Indian Ocean

The response variable for all analyses is the total count of incidences within each cell of the data matrix. For the full data matrix, therefore, there is a total count for each ICD9-month-branch-grade-platform-location combination. For consistency, divisions of the independent variables are called classes and divisions of ancillary variables called groups.

2.3 Preliminary Calculations

The original data set is made up of all outpatient encounter records between 1 January 1989 and 30 June 1989 and all Inpatient encounter records between 1 January 1980 and 31 December 1984. The data set contains 69,637 records. For purposes of the present analysis, these records are error checked and the information contained therein reduced to a simple matrix format. Error-checking consists of reading each record; extracting the date, service branch, pay grade, platform, location, and ICD9-CM categories reported; and then range-checking all codes. Any record with one or more errors is rejected in its entirety. In addition, only first visit records are retained.

If the original data record contains more than one ICD9-CM category entry, each entry is considered separately. In other words, if an original data record contains three ICD9-CM entries, it is treated as three separate records with different ICD9-CM codes but with the same date, service branch, pay grade, platform, and location. Each record of this type is defined as one incidence of the appropriate ICD9-CM category. Each incidence is summed into six individual matrices. The axes for the matrices are as follows:

1. multivariate ICD9-CM categories vs. month of the year;
2. univariate ICD9-CM categories vs. month of the year;
3. multivariate ICD9-CM categories vs. ancillary variables;
4. month of the year vs. ancillary variables;
5. date vs. univariate ICD9-CM categories;
6. date vs. multivariate ICD9-CM categories.

These six reduced data matrices support all subsequent analyses and sample size allocations. Their uses and structure are described in the remaining sections of chapter 2 and chapter 3.

2.4 Classification Analysis

The classification analysis uses matrix 1 (multivariate ICD9-CM categories vs. month) to place ICD9-CM categories and months into internally homogeneous classes. The classes are chosen so that patterns of incidence rates within each class are consistent. Consistency means ICD9-CM categories in a given class have similar patterns of incidence rates over all months and vice versa. For example, consider the following three categories over six months:

	Jan	Feb	Mar	Apr	May	Jun
A	500	200	100	800	100	300
B	600	300	200	900	200	400
C	100	200	300	400	500	600

Categories A and B have similar patterns since B = A + 100; they would, therefore, be placed in the same class. Category C is dissimilar in pattern to both A and B and would, therefore, not to be classed with either.

For classification analysis, the input matrix is of the form:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{1}$$

where n is the number of multivariate ICD9-CM categories (29) and m is the number of months of the year (12). Two analyses are carried out: one on ICD9-CM categories and one on months. The ICD9-CM analysis is discussed below. The analysis on months is exactly analogous with the subscripts transposed.

The first step is to calculate from matrix X an n x n similarity matrix:

$$S = \begin{vmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{vmatrix} \tag{2}$$

where

$$s_{ij} = \frac{1}{m} \sum_{k=1}^{m} \frac{\min(x_{ik}, x_{jk})}{\max(x_{ik}, x_{jk})} \tag{3}$$

All of the elements of matrix S are in the range [0,1]; n = 29.

The next step is cluster, classification, and construction of a binary tree dendrogram. Clustering involves ordering the ICD9-CM categories so that similar categories are close together and dissimilar categories are farther apart. Clustering is accomplished by an iterative process on matrix S. First, a correlation coefficient and the probability that the coefficient is not equal to zero is calculated for all pairs of ICD9-CM categories and the largest probability identified. The ICD9-CM categories with the highest probability are placed adjacent to one another in the cluster order. If the probability is ≥ 90%, the ICD9-CM categories are also placed together in the same class. Matrix S is then reduced to an n-1-by-n-1 matrix by averaging the rows corresponding to the chosen categories and then replacing those two rows with a single row consisting of the averages. The same averaging and replacement process is carried out on the two columns corresponding to the chosen ICD9-CM categories. Thus, once two ICD9-CM categories are paired, the pair is treated as a single entity throughout the remainder of the clustering-reduction process.

The clustering-reduction process is continued until matrix S is reduced to a single value, indicating that all ICD9-CM categories are ordered and are placed in their appropriate classes. The entire process is then repeated beginning with equation (2) based on months of the year. Note here that n becomes the number of months. Two pieces of information from the clustering process are retained: the order of the ICD9-CM categories and the probabilities associated with their correlation coefficients. This information is used to build the binary tree dendrograms. In addition, the membership of each class of ICD9-CM categories and months is retained. The only information that is directy used in subsequent analyses is the cluster order and class membership. However, the dendrogram provides a superb visual representation of the relationship among ICD9-CM categories or months. Examples of dendrograms can be found in section 4.1.

A visual representation of relations among variables is called the two-way table. For example, in the original data matrix X, rows and columns are arranged more or less arbitrarily and, therefore, do not exhibit definite patterns. A two-way table is used to expose otherwise hidden associations. To construct the two-way table, matrix X is rearranged so both ICD9-CM categories and months are in cluster order. Then each element of X is replaced with its relative magnitude scaled between zero and one. For the graphical presentation, the elements of X are depicted by plotting characters that vary in optical density; the higher the relative magnitude, the denser the plotting character. In addition, lines are added separating ICD9-CM and month classes. The resultant plot shows patterns of high and low association between individual variables and classes. Three two-way tables are prepared: ICD9-CM vs. months, ICD9-CM vs. ancillary variables, and months vs. ancillary variables.

2.5 Multiple Discriminant Analysis

Multiple discriminant analysis may be conceptualized as an extension of univariate analysis of variance to multiple dependent variables. It is used to determine the extent and manner in which previously defined classes of independent variables may be differentiated by a set of ancillary variables. For instance, multiple discriminant analysis is used to determine whether service branches (USN and USMC) are statistically significant predictors of incidence rates among ICD9-CM or month classes.

The analysis proceeds as follows for ICD9-CM categories; it is exactly analogous for months. Matrices P, T, and W are formed from matrix Y for each ICD9-CM class:

$$P_{mm} = (Y_{mn})' \; Y_{mn} \tag{4}$$

$$T_m = (Y_{mn})' \; U_n \tag{5}$$

$$W_{mm} = P_{mm} - \left(\frac{1}{n}\right) \; [T_m \; (T_m)'] \tag{6}$$

where Y is a matrix of incidence rates with n ICD9-CM categories in the class being treated and m is the number of ancillary variables (e.g., m = 2 for service branch or m = 14 for location). P is a matrix of raw cross products, T is a matrix of raw sums, and W is a matrix of deviation cross products. U is a unit vector of length n.

P, T, W, and n are accumulated over all classes; henceforth, these matrices refer to the accumulated matrices. The following matrices are then formed:

$$C_{mm} = \left(\frac{1}{n}\right) \left[P_{mm} - \left(\frac{1}{n}\right) T_m \; (T_m)'\right] \tag{7}$$

$$A_{mm} = \left(\frac{1}{n}\right) C_{mm} - W_{mm} \tag{8}$$

where C is the within-class variance-covariance matrix and A is the among-class variance-covariance matrix.

Finally,

$$F_{mm} = (W_{mm})^{-1} A_{MM} \tag{9}$$

The next step is to extract the eigenvectors and values of the asymmetrical product matrix F. The vector of eigenvalues of F is denoted E and the eigenvalue matrix as V. Vector E is analogous to an F-ratio in single classification analysis of variance, each element being the ratio of within-class and among-class variation for the appropriate variable. The columns of V are analogous to factor dimensions; they represent independent axes defining a k-space, where k is the number of classes. The nature of the k-space is such that separation among classes is maximized. The centroid of each class is plotted in discriminant k-space by the following normalization and transformation:

$$B_{mk} = V_{mk} \, (E_{\Delta k})^{-\frac{1}{2}} \tag{10}$$

$$D_{nk} = Y_{nm} \, B_{mk} \tag{11}$$

where each row of D contains the k transformed values for the corresponding independent variable. The centroids for independent variable classes are found by averaging the columns of D over each class, creating class centroids for each class, one for each of the k discriminant dimensions. The process also produces ancillary variable means and variances.

To obtain ancillary variable groups on which to base sampling strata, an F-ratio test is performed on each ancillary variable. The numerator for each F-ratio is the square of the corresponding diagonal element of matrix F divided by the denominator degrees of freedom. Denominator degrees of freedom are c-1, where c is the number of independent variable classes. The numerator of the F-ratio is the sum of squares of the remaining elements in the corresponding row of matrix F divided by the numerator degrees of freedom. Numerator degrees of freedom are t-m, where t is the number of independent variables and m is the number of months. Each F-ratio is tested for statistical significance by calculating the probability that the ratio (with c-1 and t-m degrees of freedom) is greater than 1.0.

All ancillary variables found to be significant are submitted to Ryan-Einot-Grabriel-Walsh (REGW) and Tukey-Kramer (TK) multiple range tests. The purpose is to determine which ancillary variables are not significantly different from one another and can, therefore, be grouped together. Only significantly different ancillary variables are considered because both tests are consistent with the above F-test; only significant ancillary variables can form significantly different groups.

Both tests are provided since they use different logic to arrive at their groupings. In both cases, mean incidence rates (averaged over all independent variable classes) are calculated for each significant ancillary variable and placed in ascending order; overall variance of the means (for all

classes) is also calculated. Here the two methods diverge.

REGW is a step-down method; i.e., it compares means beginning with the most different and proceeding to the least different. For example, if there are k means, it first tests the set of means {1,...,k}. If this set is found to be homogeneous, it stops and declares that all means in the set belong to the same group. If, on the other hand, REGW finds that the set {1,...,k} is heterogeneous, it will test the set {1,...,k-1}, etc., until it finds the largest homogeneous set. The procedure continues until all significant ancillary variables are grouped. Homogeneity of means for set {i,...,j} is rejected if

$$\frac{n}{(p-1)s^2}\left[\sum_{k=i}^{j}\overline{Y_k} - \frac{1}{k}\left(\sum_{k=i}^{j}\overline{Y_k}\right)^2\right] \geq F(\gamma_p; p-1, \nu) \qquad (12)$$

where

$$\gamma_p = 1 - (1-\alpha)^{\frac{p}{k}} \qquad p < k - 1 \qquad (13)$$

and

$$\gamma_p = \alpha \qquad P \geq k - 1 \qquad (14)$$

KT is a pair-wise comparison method; i.e., it compares all pairs of means in order to arrive at the ancillary variable groupings. Final groups are defined as those with overlapping homogeneous means. For instance, if means 1 and 2, and 2 and 3 are homogeneous, then the set {1,2,3} is also homogeneous. Two ancillary variables are considered significantly different if

$$\frac{\left|\overline{Y_i} - \overline{Y_j}\right|}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \geq M(\alpha; c, \nu) \qquad (15)$$

where $m(\alpha;c,\nu)$ is the α level critical value of a studentized maximum modulus distribution with c independent normal random variables with ν degrees of freedom and $c = k(k-1)/2$.

If a discrepancy in results is encountered, the REGW results are used to establish sampling strata. Each homogeneous group of ancillary variables is defined as a sampling stratum. All ancillary variables not found significant during the discriminant analysis, and thus not included in the multiple range comparisons, are placed in a special purpose stratum of their own.

2.6 Time Series Analysis

The time series analysis consists of creating periodograms or power spectral graphs based on one of two different data sets. For multivariate time series, the periodogram is based on total incidence rate of all ICD9-CM categories for each day from 1 January 1980 through 31 December 1984 (1,827 days). For univariate time series, individual periodograms for each of the 29 major ICD9-CM categories is calculated over the same time period.

The method of calculation is called the maximum entropy method. It is based on the autocorrelation ϕ_j which is the average of all lags of length j. Given a data set from x_1 through x_N, one estimate is

$$\Phi_j = \Phi_{-j} \approx \frac{1}{N+1-j} \sum_{i=0}^{N-j} x_i x_{i+j} \qquad j = 0, 1, 2, \ldots, N \tag{16}$$

The Fourier transform of the autocorrelaton is equal to the power spectrum; i.e., simply a Laurent series:

$$P \approx \frac{a_0}{\left| 1 + \sum_{k=1}^{M} a_k z^k \right|^2} \tag{17}$$

where the coefficients of equation (17) are satisfied by

$$\frac{a_0}{\left| 1 + \sum_{k=1}^{M} a_k z^k \right|^2} \approx \sum_{j=iM}^{M} \Phi_j z^j \tag{18}$$

where M is the number of coefficients included in the solution. The coefficients of equation (18) in turn satisfy the following symmetrical Toerplitz matrix:

$$\begin{bmatrix} \Phi_0 & \Phi_1 & \Phi_2 & \cdots & \Phi_M \\ \Phi_1 & \Phi_0 & \Phi_1 & \cdots & \Phi_{M-1} \\ \Phi_2 & \Phi_1 & \Phi_0 & \cdots & \Phi_{M-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Phi_M & \Phi_{M-1} & \Phi_{M-2} & \cdots & \Phi_0 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} a_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{19}$$

which is solved for the a_j. Equation (17) is then plotted. Fundamental periods are obtained by locating the period with maximum power. This is done independently for each periodogram calculated.

2.7 Multivariate Sample Size Allocation

This process is entirely straightforward. The basis for determining the number of patient encounters required to adequately detect ratios ≥ 2.0 between units is stratified random sampling. Given that the strata are set, two pieces of information are required to complete the sample size allocation for each stratum: relative stratum size and within-stratum variances. Both pieces of information are available at this point in the analysis.

Sampling strata are based on the ancillary variable groups defined during the multiple discriminant analysis. The size of each stratum is taken as proportional to the total DNBI incidence rate within each stratum. Within stratum variances are based on ICD9-CM and month classes within each strata.

Total sample size for all strata is calculated as

$$n = \frac{\left(\sum_{i=1}^{h} \frac{N_i S_i}{N}\right)^2}{V + \sum_{i=1}^{h} N_i S_i} \tag{20}$$

where N_i and S_i are the number and standard deviation of DNBI incidences in stratum i, N is total N_i, and V is total allowable variability calculated as

$$V = \frac{N\, d^2}{t^2_{N,1-\alpha}} \tag{21}$$

where d = 2.0, the allowed deviation, and t is student's t with N degrees of freedom and probability $1-\alpha$.

Before equation (20) is applied, the stratum standard deviations S_i are adjusted for the length of the fundamental period calculated by the time series analysis. For this adjustment, it is assumed that the primary source of variability over time occurs at the fundamental period. In addition, the source of time variability in reduced historical data matrices has a period of one month, thus smoothing shorter period variability. Therefore, the S_i are multiplied by the ratio 30/P where P is the fundamental period. Finally, the sample size for each stratum is calculated as

$$N_i = N \frac{N_i S_i}{\sum_{i=1}^{h} N_i S_i} \tag{22}$$

The above process is carried out separately for each of the four ancillary variables. Total sample size for all four analyses is adjusted to the maximum value of N for any of the four. Therefore, all allocations have the same values for N.

2.8 Univariate Sample Size Allocation

Univariate sampling involves sampling for a single ICD9-CM category. Sample size allocation is mostly identical to multivariate allocation. However, there are two essential differences. First, since the analysis uses the 130 detailed ICD9-CM categories one at a time, within- stratum variability is calculated based only on the class in which the ICD9-CM category is found. Second, the manner in which the estimated sample size is applied is different. Here, the procedure is to continue sampling within the stratum until the specified number (N_i) of encounters for the specified ICD9-CM category occurs.

3.0 DESCRIPTION OF COMPUTER PROGRAM

3.1 Overview

The computer system developed as part of this project contains all of the analyses and outputs described in chapter 2. The data reductions (section 2.3) are programmed in ANSI Fortran and compiled and run of the NHRC VAX-11785. These programs involve manipulating very large data files and reducing the information to much smaller intermediate data files. The intermediate files are transferred to an IBM-PC where the remainder of the analyses are run.

The classification, multiple discriminant, time series, and sample allocation analyses are programmed in Fortran and are designed to run on an IBM-PC with an 80286 or 80287 processor and math coprocessor. The system is currently configured for EGA or VGA graphics. The overall system is controlled by a compiled program written in FORCE 2.0, an integrated control, menu, and database management language. The system is completely menu-driven.

During the course of the analysis, information is passed from program to program via temporary data files. Before any output can be requested, a complete analysis is run and all output is generated for quick and immediate retrieval. Both screen and printer hard-copy output is available for all analyses.

As mentioned above, the entire analytical system is menu-driven. Upon entering the program, the menu options are as follows:

DNBI Main Options Menu

A. Type I Error
B. Output Mode
C. Preliminary Analysis
D. Classification Results
E. Two-Way Tables
F. Multidiscriminant Results
G. Sample Stratification
Q. Quit

An option is selected by moving the highlight with the cursor keys and then pressing RETURN or by pressing the letter (A through Q) corresponding to the option. The following sections describe each of the main options.

3.2 Type I Error Option

This option is an analysis set up option. It gives the user the opportunity to set the confidence level for all subsequent analyses. The confidence level is currently limited to values between 90% and 99%. Suggested values are 90%, 95%, and 99%; the reason being that the critical value tables for the Tukey-Kramer multiple range test (section 2.5) are only available for 90%, 95%, and 99% levels. If this option is not chosen, the default value is 90%.

3.3 Output Mode Option

This option is used to set the destination for all output. The options are to send the output to the screen or to send it to the printer. The default condition is output to the screen.

3.4 Preliminary Analysis

This option runs a complete analysis and generates all output. The analysis is completely automatic; however, progress is shown by messages that appear on the screen. A certain amount of duplication exists in each of the analytical modules; this is necessary because each module is designed to operate either in concert with the others or independently. Throughout the following discussion, the file extensions ROW and COL refer to ICD9-CM categories and months, respectively.

The first step in the analysis is to detect for the presence of current time series analysis results; these are contained in the file SPECTRUM.TMP. If the results are not present, then SPECTRUM.EXE is executed without time series and periodogram screen output. The second step is the classification analysis. This module contains two programs. CLSTPRP.EXE performs the actual classification and produces the documentation file. CLSTPRN.EXE defines the classes and then produces printable ASCII files containing the dendrograms. Module structure is indicated below, followed by module flow:

Program Structure - Classification Analysis Module

Program	CL Option*	Input	Output
CLSTPRP	[ROW/COL]	CLSTMAT.DAT	CLUSTER.TMP CLSTDOC.[ROW/COL] DENDRO .[ROW/COL]
CLSTPRN	[ROW/COL][PROB]	DENDRO.[ROW/COL]	CLSTDEN.[ROW/COL] -CLUSTER.TMP

* Command Line Option / designates an option

Program Flow - Classification Analysis Module

```
CLSTPRP [ROW]
COPY CLUSTER.TMP -> CLUSTER.ROW
CLSTPRN [PROB]

CLSTPRP [ROW]
COPY CLUSTER.TMP -> CLUSTER.COL
CLSTPRN [PROB]
```

The third step is the multiple discriminant analysis. The analysis is performed in eight parts and produces eight sets of output: one for each of the four ancillary variables based on ICD9-CM categories and months. There are four programs in the module. DISCPRP1.EXE and D-ISCPRP2.EXE perform a much simplified form of classification; the results are the same as those obtained from the classification module. DISCRIM.EXE performs the actual discriminant analysis and produces the output. DSTRATA.EXE performs the multiple range tests. Module structure and flow are as follows:

Program Structure - Discriminant Analysis Module

Program	CL Option	Input	Output
DISCPRP1	[ROW/COL][ANC]	CLSTMAT.DAT ANCVAR .[ROW/COL]	DISCRIM.TMP
DISCPRP2	[ROW/COL][PROB]	DISCRIM.TMP DENDRO .[ROW/COL]	DISCRIM.TMP
DISCRIM	[PROB]	DISCRIM.TMP	DISCRIM.TMP MDOC .TMP MPLT .TMP
DSTRATA	[PROB]	DISCRIM.TMP STDRNGE.DAT	MSTR .TMP VSTR .TMP

Program Flow - Discriminant Analysis Module

Repeat for rcopt = ROW,COL		
Repeat for avopt = 1,2,3,4 (ancillary vars)		
DISCPRP1 [rcopt] [avopt]		
DISCPRP2 [rcopt] [PROB]		
COPY DISCRIM.TMP	->	DISCRIM.[rcopt]
DISCRIM [PROB]		
DSTRATA [PROB]		
COPY MDOC.TMP	->	MDOC[avopt].[rcopt]
COPY MSTR.TMP	->	MSTR[avopt].[rcopt]
COPY MPLT.TMP	->	MPLT[avopt].[rcopt]
COPY VSTR.TMP	->	VSTR[avopt].[rcopt]

The fourth step is the construction of the two-way tables. This module produces three tables: ICD9-CM vs. months, ICD9-CM vs. ancillary variables, and months vs. ancillary variables. The module contains five programs. TWTPRP1.EXE and TWTPRP2.EXE perform a simplified form of classification. TWTPRP3.EXE prepares ancillary variable information for incorporation into the appropriate two-way tables. TWTPRN1.EXE constructs the ICD9-CM vs. month table. TWTPRN2.EXE constructs the ICD9-CM vs. ancillary variable and month vs. ancillary variable tables. Module structure and flow are as follows:

Program Structure - Two-Way Table Module

Program	CL Option	Input	Output
TWTPRP1	[ROW/COL]	CLSTMAT.DAT	DENDRO.TMP TWT .TMP
TWTPRP2	[ROW/COL][PROB]	DENDRO.TMP TWT .TMP	TWT .[ROW/COL]
TWTPRP3	[ROW/COL]	ANCVAR.[ROW/COL]	TWTTMP.ANC
TWTPRN1	(none)	TWTTMP. ROW TWTTMP. COL CLSTMAT.DAT	TWOWAY.RC
TWTPRN2	[ROW/COL]	TWTTMP.[ROW/COL] ANCVAR.[ROW/COL] TWTTMP.ANC	TWOWAY.[RA/CA]

Program Flow - Two-Way Table Module

TWTPRP1 [ROW]
TWTPRP2 [ROW] [PROB]
TWTPRP1 [COL]
TWTPRP2 [COL] [PROB]
TWTPRN1

TWTPRP3 [ROW]
COPY TWTTMP.ANC -> TWTTMP.RA
TWTPRN2 [ROW]

TWTPRP3 [COL]
COPY TWTTMP.ANC -> TWTTMP.CA
TWTPRN2 [COL]

The final step in the preliminary analysis is stratification and sample size allocation. This module includes three programs. VSTRPRP.EXE and USTRATA.EXE define the sampling strata based on the ancillary variable groups for multivariate and univariate sampling, respectively. VSTRATA.EXE performs the multivariate sample size allocation and produces output files. Module structure and flow is given below:

Program Structure - Sample Sizw Allocation Module

Program	CL Option	Input	Output
VSTRPRP	(none)	ANCTVAR.COL VSTR* .COL	VSTRATA1.TMP VSTRATA2.OUT
USTRPRP	(none)	ANCTVAT.ROW VSTR* .ROW	USTRATA1.TMP USTRATA2.OUT
VSTRATA	(none)	ANCTVAR .COL VSTRATA1.TMP	USTRATA1.OUT

* indicates all possible values

Program Flow - Sample Allocation Module

VSTRPRP
VSTRATA
VSTRPRP

3.5 Classification Results

This module sends classification documentation and dendrogram output to either the screen or the printer, according to which output mode is selected (section 3.3). If printed output is chosen, documentation and dendrograms for ICD9-CM categories, months, or both can be printed. If screen output is chosen, results for either ICD9-CM or months can be viewed separately. Module structure is as follows:

Program Structure - Classification Results Module

Program	CL Option	Input	Output
FLIST	(n/a)	CLSTDOC.[ROW/COL]	SCREEN
COPY	(n/a)	CLSTDOC.[ROW/COL]	PRINTER
CLSTPLT	[ROW/COL][PROB]	DISCRIM.[ROW/COL] CLUSTER.[ROW/COL] DENDRO .[ROW/COL]	SCREEN
COPY	(n/a)	CLSTDEN.[ROW/COL]	PRINTER

3.6 Multiple Discriminant Results

This module sends multiple discriminant analysis results to either the screen or the printer. ICD9-CM or months can be specified as well as ancillary variables (service branch, platform, location, or paygrade). Module structure is given below:

Program Structure - Discriminant Results Module

Program	CL Option	Input	Output
FLIST	(n/a)	MDOC*.[ROW/COL] MSTR*.[ROW/COL]	SCREEN
COPY	(n/a)	MDOC*.[ROW/COL] MSTR*.[ROW/COL]	PRINTER
DISCPLT	[rc] [av]	MPLT[av].[rc]	SCREEN

3.7 Two-Way Table Results

This module sends two-way table results to either the screen or the printer. Three tables are available for output: ICD9-CM vs. months, ICD9-CM vs. ancillary variables, and months vs. ancillary variables. Module structure is as follows:

Program Structure - Two-Way Table Results Module

Program	CL Option	Input	Output
COPY	(n/a)	TWOWAY.RC	PRINTER
COPY	(n/a)	TWOWAY.RA	PRINTER
COPY	(n/a)	TWOWAY.CA	PRINTER
TWTPLT1	(none)	CLSTMAT.DAT TWTTMP .ROW TWTTMP .COL	SCREEN
TWTPLT2	[ROW/COL]	CLSTMAT.DAT TWTTMP .[ROW/COL] ANCVAR .[ROW/COL]	SCREEN

3.8 Stratification and Sample Allocation Results

The module performs two tasks. First, it sends multivariate and univariate stratification and multivariate sample allocation to the screen or printer. In addition, it controls the calculation and presentation of univariate sample allocations. When univariate sample allocation is chosen, control is turned over to UNIVSTR.EXE which is used to select an ICD9-CM category for allocation and then to run USTRATA.EXE.

Program Structure - Allocation Results Module

Program	CL Option	Input	Output
COPY	(n/a)	VSTRATA1.OUT	PRINTER
COPY	(n/a)	VSTRATA2.OUT	PRINTER
COPY	(n/a)	USTRATA1.OUT	PRINTER
COPY	(n/a)	USTRATA2.OUT	PRINTER
FLIST	(n/a)	VSTRATA1.OUT	SCREEN
FLIST	(n/a)	VSTRATA2.OUT	SCREEN
FLIST	(n/a)	USTRATA1.OUT	SCREEN
FLIST	(n/a)	USTRATA2.OUT	SCREEN
UNIVSTR	(none)	(none)	USTRATA1.OUT
USTRATA	[ICD9 #]	TWTTMP .ROW ICDCATE .DAT UNIVMAT .DAT USTRATA1.TMP	USTRATA1.OUT

3.9 Time Series Analysis

This module consists of two programs, both of which send their output to the screen. There is no printer output for this module. TIMESER.EXE plots DNBI incidence rate against time and SPECTMEM.EXE calculates and plots the spectral power density and plots relative magnitude against period. The structure of this module follows:

Program Structure - Time Series Module

Program	CL Option	Input	Output
TIMESER	(none)	SPECTRUM.DAT	SCREEN
SPECTMEM	(none)	SPECTRUM.DAT	SPECTRUM.TMP SCREEN

3.10 Data File Cross Reference

This section lists and describes the contents of all of the major data files in the analytical and output system. Initial files are files derived from historical data; these are not changed. Intermediate files are created by the system to communicate information between programs. Output files are ASCII files for screen or printer output.

File	Description
	INITIAL FILES
CLSTMAT.DAT	counts in ICD9-CM (29) x months matrix
UNIVMAT.DAT	counts in ICD9-CM (130) vs. months matrix
ANCVAR .[ROW/COL]	counts for set of 29 ICD9-CM (ROW) or month (COL) vs. ancillary variables
ANCTVAR.[ROW/COL]	counts for set of 130 ICD9-CM (ROW) or month (COL) vs. ancillary variables
SPECTRUM.DAT	days since 1 Jan 80 and total DNBI counts
STDTRNGE.DAT	table of studentized range critical values
	INTERMEDIATE FILES
CLUSTER.[ROW/COL]	ICD9-CM or month codes in cluster order
DENDRO .[ROW/COL]	ICD9-CM or month dendrogram linkage data
DISCRIM.[ROW/COL]	ICD9-CM or month classes and ancillary group data
VSTR[A].[ROW/COL]	ICD9-CM or month vs. ancillary variable "A" ancillary group definitions
MPLT[A].[ROW/COL]	ICD9-CM or month vs. ancillary variable "A" class data transformed to discriminant axes
TWTTMP .[ROW/COL]	ICD9-CM or month variables and classes in two-way table order
	OUTPUT FILES
CLSTDOC.[ROW/COL]	ICD9-CM or month classification
CLSTDEN.[ROW/COL]	ICD9-CM or month dendrograms
MDOC[A].[ROW/COL]	ICD9-CM or month vs. ancillary variable "A" discriminant documentation
MSTR[A].[ROW/COL]	ICD9-CM or month vs. ancillary variable "A" multiple range test results
TWOWAY .RC	ICD9-CM vs. month two-way table
TWOWAY .RA	ICD9-CM vs. ancillary two-way table
TWOWAY .CA	month vs. ancillary two-way table
VSTRATA1.OUT	multivariate sample strata
VSTRATA2.OUT	multivariate sample allocations
USTRATA1.OUT	univariate sample strata
USTRATA2.OUT	univariate sample allocations

4.0 RESULTS

4.1 Preliminary Data Reduction

Preliminary data reduction was performed on inpatient encounter records from 1 January 1980 through 31 December 1984 and outpatient records from 1 January 1989 through 30 June 1989. The files represent 1,827 days of data covering all 29 major DNBI reporting categories. Table 4 is a summary of the data reduction and error-checking.

TABLE 4. File Preparation Report

		Outpatient	Inpatient	Combined
Totals:				
	# Records	23,648	34,712	58,360
	# Cases	24,733	44,904	69,637
Cases Rejected:				
	ICD9-CM	971	1,386	2,357
	Date	0	0	0
	Branch	29	0	29
	Platform	12	687	699
	Location	4	18,183	18,187
	Pay grade	252	0	252

4.2 Classification Analysis

Figure 1 is the dendrogram resulting from the classification of the 29 major ICD9-CM categories. The darker lines within the body of the dendrogram represent linkages for which the probability that the entities linked belong to the same class is 95% or greater. Note that 16 classes contain from one to three ICD9-CM categories; this is a significant reduction over the original 29 major categories. Figure 2 shows the linkages along with the similarity coefficient, the probability the linkage is not due to chance alone (CCCS), and the upper and lower bounds on the probability.

Figure 3 is the dendrogram for the classification of months of the year. Note that the number of classes is only six, again a significant reduction. Figure 4 for months is analogous to Figure 2. Figures 2 and 4 are available for technical evaluation in case of any failure in the classification algorithms.

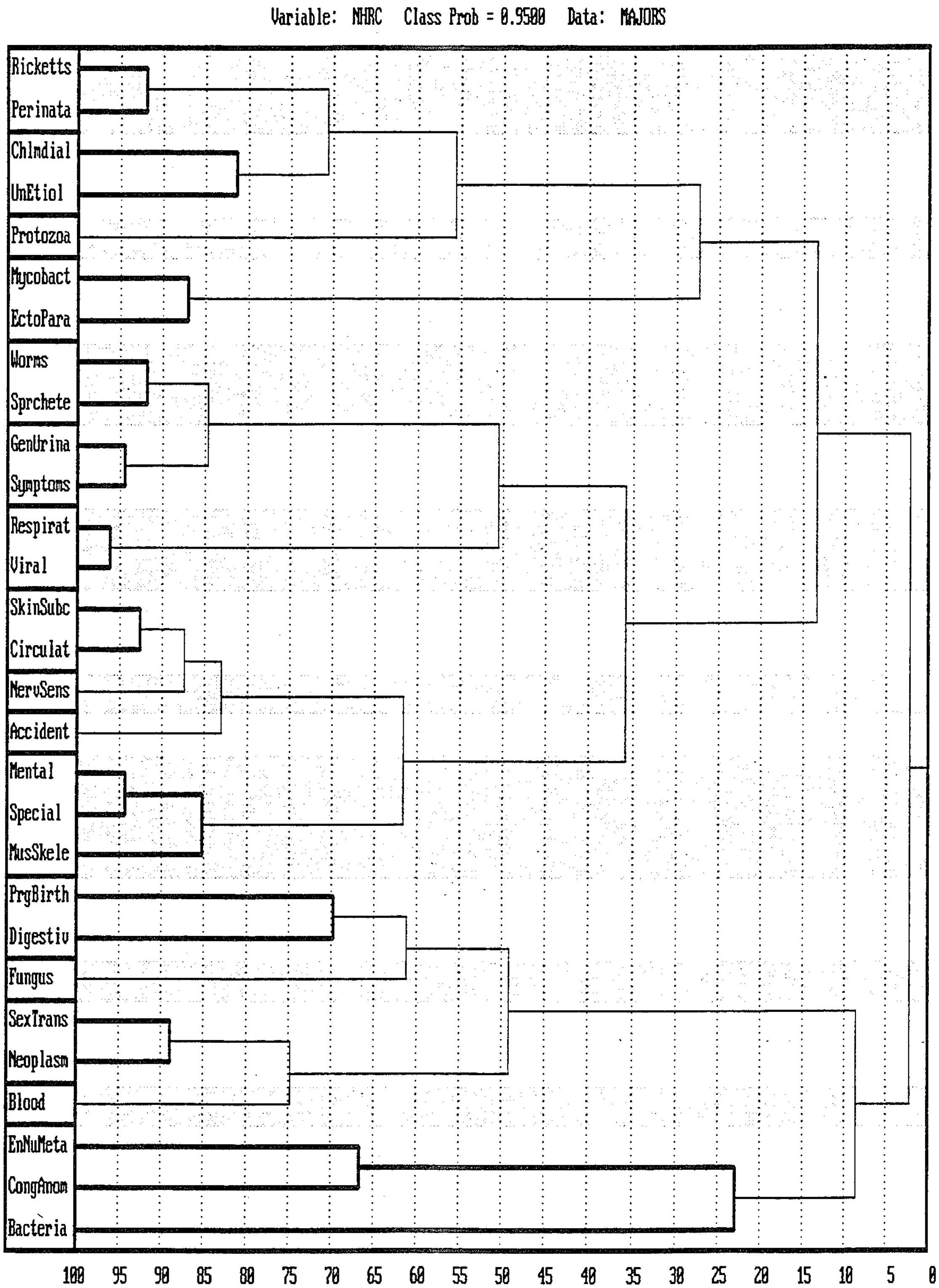

FIGURE 1. Dendrogram for ICD9-CM major categories with months as the independent variable. Major categories are divided into classes by darker horizontal lines. Class probabilities = 95%

MULTIVARIATE CLUSTER ANALYSIS

All Majors

Classification Variable : NHRCCODE
Independent Variable : MONTH
Similarity Coefficient : Pearson

Classificatin Variables: 29
Independent Variables : 12

CYCLE	LINKED		COEF	CCCS	ZL	ZU
1	2	25	0.9200	1.0000	1.0000	1.0000
1	3	9	0.8894	1.0000	1.0000	1.0000
1	5	26	0.6667	1.0000	1.0000	1.0000
1	6	7	0.8713	1.0000	1.0000	1.0000
1	8	11	0.6986	1.0000	1.0000	1.0000
1	14	15	0.8143	1.0000	1.0000	1.0000
1	16	28	0.9624	1.0000	1.0000	1.0000
1	17	18	0.9264	1.0000	1.0000	1.0000
1	19	27	0.9443	1.0000	1.0000	1.0000
1	20	22	0.9187	1.0000	1.0000	1.0000
1	24	29	0.9453	1.0000	1.0000	1.0000
2	2	14	0.7075	0.7976	0.0000	0.9769
2	3	12	0.7484	0.9417	0.9417	0.9417
2	17	23	0.8739	0.6713	0.6713	0.6713
2	19	21	0.8527	0.9884	0.9884	0.9884
2	20	24	0.8467	0.8955	0.3070	0.9886
3	1	17	0.8307	0.7317	0.0000	0.9683
3	8	10	0.6126	0.8133	0.8133	0.8133
4	1	19	0.6166	0.9271	0.8259	0.9704
4	3	8	0.4923	0.7737	0.4331	0.9210
4	2	13	0.5572	0.8126	0.3747	0.9541
4	4	5	0.2292	0.9707	0.9707	0.9707
5	16	20	0.5056	0.9848	0.9536	0.9951
6	1	16	0.3580	0.8325	0.7486	0.8901
7	2	6	0.2739	0.9243	0.8197	0.9692
8	1	2	0.1337	0.8768	0.8392	0.9061
8	3	4	0.0856	0.9441	0.8923	0.9713
9	1	3	0.0219	0.9301	0.9157	0.9421

FIGURE 2. Backup documentation for dendrogram of ICD9-CM categories. COEF is the correlation coefficient from matrix S. CCCS, ZL, and ZU are the probability that COEF $\neq 0.0$ and the lower and upper confidence limits, respectively, of the probability.

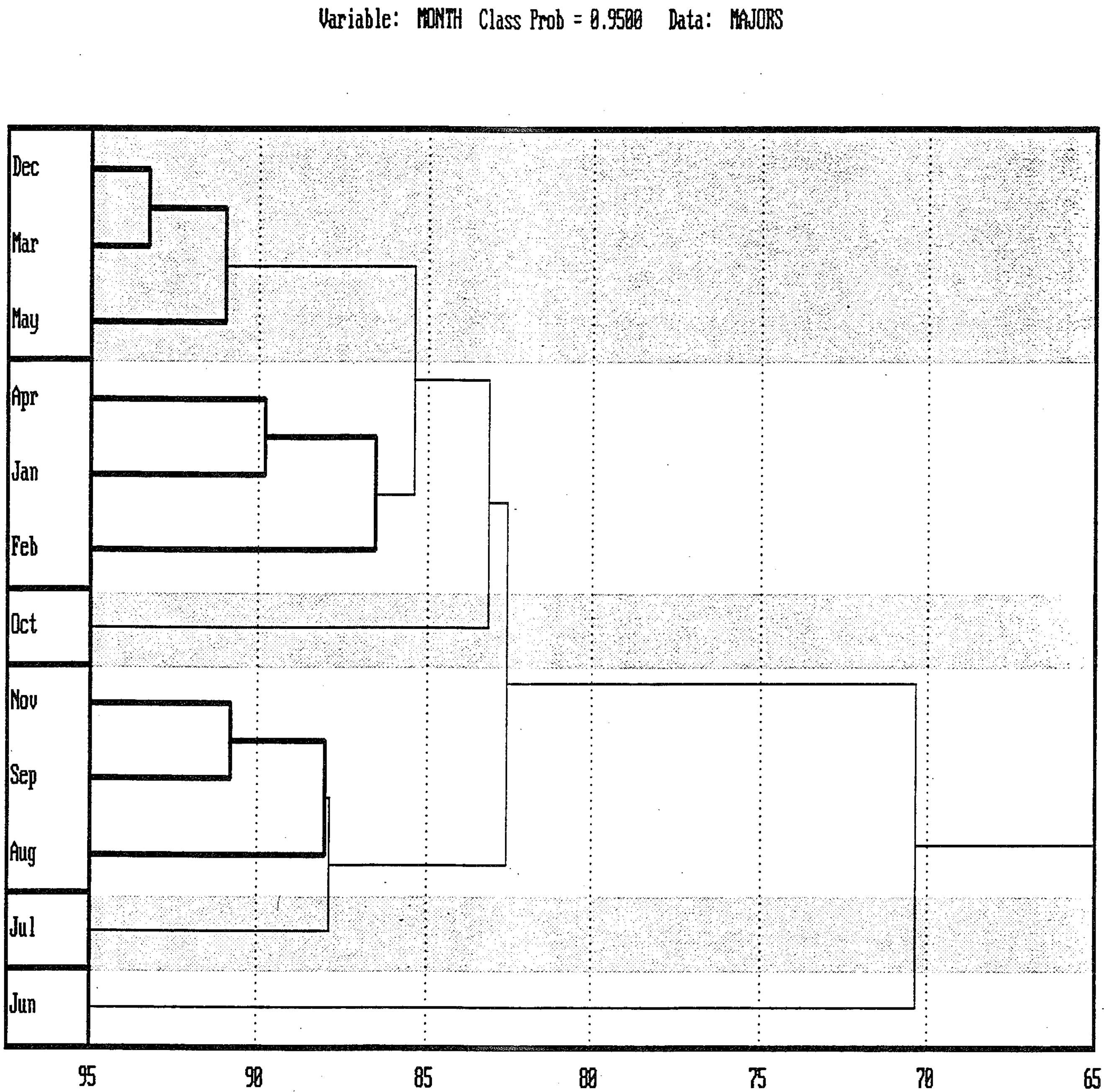

FIGURE 3. Dendrogram for months with ICD9-CM categories as the independent variable. Variable list is divided into classes by darker horizontal lines. Class separation probabilities = 95%.

MULTIVARIATE CLUSTER ANALYSIS

All Majors

Classification Variable : MONTH
Independent Variable : NHRCCODE
Similarity Coefficient : Pearson

Classificatin Variables: 12
Independent Variables : 29

CYCLE	LINKED		COEF	CCCS	ZL	ZU
1	1	4	0.9081	1.0000	1.0000	1.0000
1	6	7	0.9329	1.0000	1.0000	1.0000
1	9	11	0.8981	1.0000	1.0000	1.0000
2	6	8	0.9099	1.0000	1.0000	1.0000
2	9	10	0.8653	0.9781	0.9781	0.9781
3	1	5	0.8798	0.9641	0.9641	0.9641
4	1	3	0.8783	0.8302	0.0571	0.9809
5	6	9	0.8540	0.9167	0.7624	0.9723
6	2	6	0.8317	0.9044	0.7754	0.9609
7	1	2	0.8261	0.6619	0.4811	0.7887
8	1	12	0.7036	0.8791	0.8093	0.9244

FIGURE 4. Backup documentation for dendrogram of months of the year. COEF is the correlation coefficient from matrix S. CCCS, ZL, and ZU are the probability that COEF ≠ 0.0 and the lower and upper confidence limits, respectively, of the probability.

4.3 Two-way Tables

Figures 5 shows an example of a two-way tables for ICD9-CM vs. age groups. The table show graphically the relationships between age and ICD9-CM incidence rates.

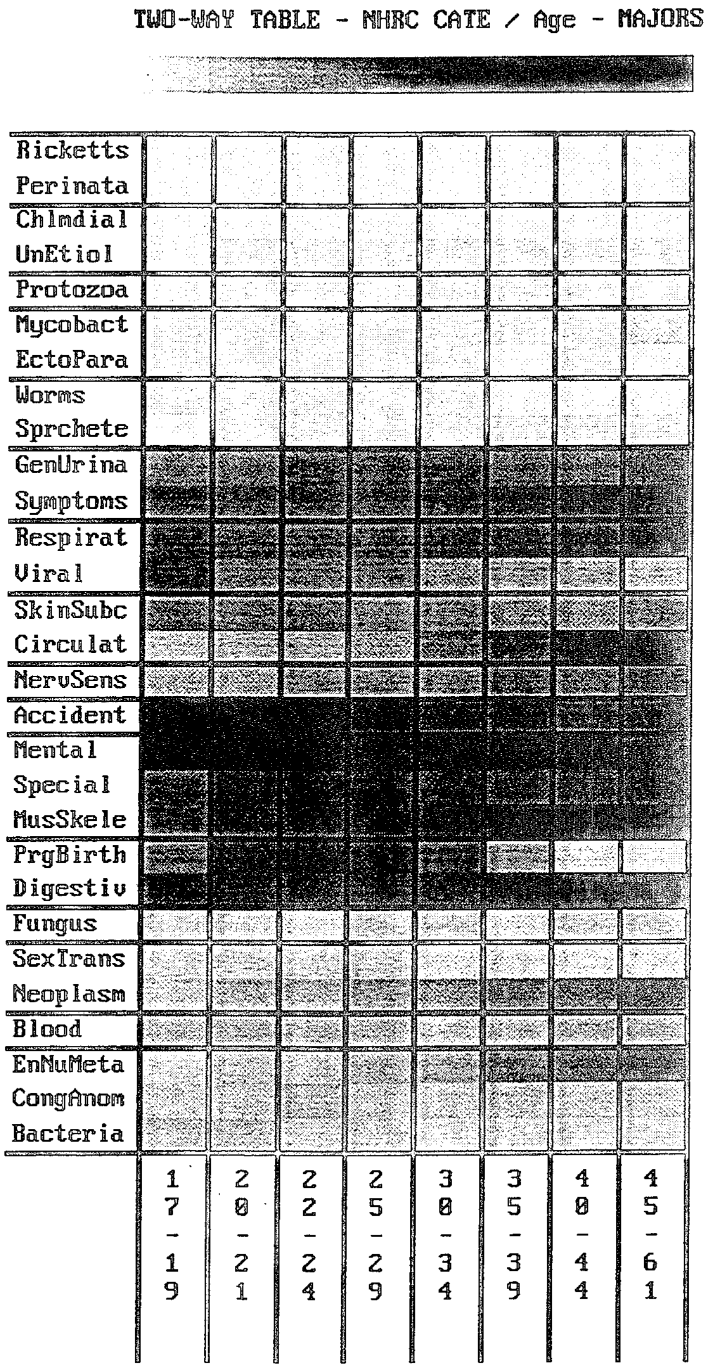

FIGURE 5. Two-way table for ICD9-CM categories vs. Age groups. Plotting characters of increasing optical density indicate $\log_{10}$ increases in DNBI incidence rates.

4.4 Multiple Discriminant Analysis

Figures 8 through 11 show the multiple discriminant results based on ICD9-CM categories and classes for service branch, platform, location, and pay grade, respectively. Figures 12 through 15 show the same based on months and month classes. In the section of each figure labeled "F-test for Ancillary Variables," those variables marked with an asterisk are significant predictors of DNBI rates within independent variable classes. Only those marked with asterisks are submitted to the multiple range tests and, therefore, are the only ones considered as potential ancillary variable groups.

Figures 16 through 19 show the results of the multiple range tests, establishing ancillary variable groups based on ICD9-CM categories and classes. Figures 20 through 23 show the establishment of ancillary variable groups based on months and month classes. A double line indicates a variable realization in a group by itself. Otherwise, groups defining sampling strata are bracketed.

Page intentional left blank

MULTIPLE DISCRIMINANT ANALYSIS:

Classified Variables (rows): ICD9-CM
Ancillary Variables (cols): BRANCH

F-TESTS FOR ANCILLARY VARIABLES:

Numerator df: 12
Denominator df: 15

Variable /	F-Ratio	Probability	
USN	2.3850	.942963	*
USMC	1.4552	.756981	

FIGURE 8. Results of multiple discriminant analysis on ICD9-CM categories with service branch as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
-------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
-------------------------------------------
Classified Variables (rows):  ICD9-CM
Ancillary  Variables (cols):  PLATFORM
-------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:      12
Denominator df:      15
```

Variable /	F-Ratio	Probability	
Ashore	1.7616	.850621	
AE	1.8709	.874282	
AFS	1.9409	.887363	
AGF	1.1936	.632776	
AO	1.4092	.738561	
AOE	1.1969	.634652	
AOR	1.5828	.801613	
AR	2.1978	.924312	*
ARS	3.5830	.988721	*
AS	2.2883	.934037	*
AVT	.7085	.276414	
CG	2.0926	.911031	*
CGN	2.6714	.962480	*
CV	2.2129	.926041	*
CVN	2.1078	.913092	*
DD	2.2494	.930038	*
DDG	1.8532	.870722	
FF	1.9231	.884169	
FFG	2.2476	.929842	*
LCC	2.3611	.940887	*
LHA	4.4019	.995446	*
LKA	1.0693	.555735	
LPD	1.7341	.843990	
LPH	1.7492	.847654	
LSD	2.3418	.939150	*
LST	1.8679	.873680	
MSO	2.3355	.938571	*

FIGURE 9. Results of multiple discriminant analysis on ICD9-CM categories with platform as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

MULTIPLE DISCRIMINANT ANALYSIS:

Classified Variables (rows): ICD9-CM
Ancillary Variables (cols): LOCATION

F-TESTS FOR ANCILLARY VARIABLES:

Numerator df: 12
Denominator df: 15

Variable /	F-Ratio	Probability	
Atl-nw	1.3285	.702885	
Car-GoM	2.6830	.963096	*
Atl-ne	2.0407	.903586	*
Med	2.3378	.938779	*
Per-Red	2.6453	.961053	*
Pac-ne	2.6769	.962775	*
Pac-cen	4.3997	.995436	*
Pac-w	2.0554	.905755	*
Pac-se	2.7442	.966171	*
Indian	1.9730	.892862	

FIGURE 10. Results of multiple discriminant analysis on ICD9-CM categories with location as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
---------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
---------------------------------------------
Classified Variables (rows):  ICD9-CM
Ancillary  Variables (cols):  PAYGRADE
---------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:      12
Denominator df:      15

Variable   /       F-Ratio     Probability
----------------------------------------
Enlisted           2.0671         .907459   *
Officer            8.2188         .999707   *
```

FIGURE 11. Results of multiple discriminant analysis on ICD9-CM categories with pay grade as the ancillary variable. * indicates statistically significant F-ratios with 1-α > 90%.

```
--------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
--------------------------------------------
Classified Variables (rows):  MONTH
Ancillary  Variables (cols):  BRANCH
--------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:        3
Denominator df:        8

Variable   /      F-Ratio     Probability
-----------------------------------------
USN               20.2303       .999240  *
USMC               5.9547       .980298  *
```

FIGURE 12. Results of multiple discriminant analysis on months with service branch as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
-------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
-------------------------------------------
Classified Variables (rows):  MONTH
Ancillary  Variables (cols):  PLATFORM
-------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:       3
Denominator df:       8
```

Variable /	F-Ratio	Probability	
Ashore	1.8206	.778939	
AE	2.0358	.812833	
AFS	1.5406	.722852	
AGF	.1678	.474745	
AO	5.8082	.978998	*
AOE	3.2527	.919256	*
AOR	1.4705	.706185	
AR	1.1642	.617629	
ARS	4.9814	.969140	*
AS	2.5702	.873135	
AVT	1.7631	.768647	
CG	2.2475	.840203	
CGN	85.4963	.999960	*
CV	12.8874	.997531	*
CVN	6.1328	.981742	*
DD	.6683	.496410	
DDG	1.3857	.684397	
FF	9.0717	.993586	*
FFG	1.7121	.759021	
LCC	.4932	.492333	
LHA	3.3541	.924208	*
LKA	.8369	.499548	
LPD	10.3529	.995526	*
LPH	3.0078	.905553	*
LSD	8.8579	.993156	*
LST	4.9755	.969051	*
MSO	.5988	.494919	

FIGURE 13. Results of multiple discriminant analysis on months with platform as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
-------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
-------------------------------------------
Classified Variables (rows):  MONTH
Ancillary  Variables (cols):  LOCATION
-------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:       3
Denominator df:       8
```

Variable /	F-Ratio	Probability	
Atl-nw	4.8304	.966736	*
Atl-ne	10.4826	.995675	*
Med	8.9758	.993398	*
Per-Red	2.7123	.884973	
Pac-ne	158.4662	.999983	*
Pac-cen	7.9389	.990791	*
Pac-w	6.2540	.982647	*
Indian	3.0843	.910123	*

FIGURE 14. Results of multiple discriminant analysis on months with location as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
-------------------------------------------
MULTIPLE DISCRIMINANT ANALYSIS:
-------------------------------------------
Classified Variables (rows):  MONTH
Ancillary  Variables (cols):  PAYGRADE
-------------------------------------------

F-TESTS FOR ANCILLARY VARIABLES:

Numerator   df:      3
Denominator df:      8

Variable   /     F-Ratio    Probability
---------------------------------------
Enlisted        17.7187        .998935  *
Officer         27.0599        .999624  *
```

FIGURE 15. Results of multiple discriminant analysis on months with pay grade as the ancillary variable. * indicates statistically significant F-ratios with $1-\alpha > 90\%$.

```
----------------------------------------
Stratification Results
----------------------------------------
Data Set:  ICD9-CM
Variable:  BRANCH
----------------------------------------

REGWF Multiple Range:     TUKEY Multiple Range:

      USN  ==                   USN  ==

     USMC  ==                  USMC  ==
```

FIGURE 16. Results of stratification on service branch based on ICD9-CM categories. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
---------------------------------------
Stratification Results
---------------------------------------
Data Set:   ICD9-CM
Variable:   PLATFORM
---------------------------------------

REGWF Multiple Range:    TUKEY Multiple Range:

         AR ==                    AR ==

        LSD -+                   LSD -+
         AS  |                    AS  |
        LCC  |                   LCC  |
        ARS  |                   ARS  |
         DD  |                    DD  |
        MSO  |                   MSO  |
        CVN  |                   CVN  |
         CV  |                    CV  |
        FFG  |                   FFG  |
         CG  |                    CG  |
        CGN  |                   CGN  |
        LHA -+                   LHA -+

        AOE -+                   AOE -+
        AOR  |                   AOR  |
        LPD  |                   LPD  |
        AVT  |                   AVT  |
        LPH  |                   LPH  |
         FF  |                    FF  |
        AGF  |                   AGF  |
        LST  |                   LST  |
        DDG  |                   DDG  |
     Ashore  |                Ashore  |
        AFS  |                   AFS  |
         AE  |                    AE  |
        LKA  |                   LKA  |
         AO -+                    AO -+
```

FIGURE 17. Results of stratification on platform based on ICD9-CM categories. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
Stratification Results

Data Set:   ICD9-CM
Variable:   LOCATION

REGWF Multiple Range:        TUKEY Multiple Range:

     Pac-se ==                    Pac-se ==

    Per-Red ==                   Per-Red ==

        Med ┐                        Med ┐
      Pac-w │                      Pac-w │
    Pac-cen │                    Pac-cen │
    Car-GoM │                    Car-GoM │
     Atl-ne │                     Atl-ne │
     Pac-ne ┘                     Pac-ne ┘

     Indian ┐                     Indian ┐
     Atl-nw ┘                     Atl-nw ┘
```

FIGURE 18. Results of stratification on location based on ICD9-CM categories. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
----------------------------------------
Stratification Results
----------------------------------------
Data Set:  ICD9-CM
Variable:  PAYGRADE
----------------------------------------

REGWF Multiple Range:     TUKEY Multiple Range:

  Enlisted ==               Enlisted ==

   Officer ==                Officer ==
```

FIGURE 19. Results of stratification on pay grade based on ICD9-CM categories. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
--------------------------------------------
Stratification Results
--------------------------------------------
Data Set:   MONTHS
Variable:   BRANCH
--------------------------------------------

REGWF Multiple Range:      TUKEY Multiple Range:

        USN ═                       USN ═

       USMC ═                      USMC ═
```

FIGURE 20. Results of stratification on service branch based on months. Strata are bracketed. A double line indicates the variable is the sole inhabitant of the stratum.

```
----------------------------------------
Stratification Results
----------------------------------------
Data Set:   MONTHS
Variable:   PLATFORM
----------------------------------------

REGWF Multiple Range:     TUKEY Multiple Range:

          AO                        AO

         LST                       LST
         ARS                       ARS
         LPH                       LPH
         AOE                       AOE
         LHA                       LHA
          FF                        FF
         CVN                       CVN
         LPD                       LPD
         CGN                       CGN
          CV                        CV
         LSD                       LSD

      Ashore                    Ashore
          AE                        AE
         AFS                       AFS
          AR                        AR
          AS                        AS
         AOR                       AOR
         AVT                       AVT
          CG                        CG
         DDG                       DDG
          DD                        DD
         LCC                       LCC
         AGF                       AGF
         FFG                       FFG
         LKA                       LKA
         MSO                       MSO
```

FIGURE 21. Results of stratification on platform based on months. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
----------------------------------------
Stratification Results
----------------------------------------
Data Set:  MONTHS
Variable:  LOCATION
----------------------------------------

REGWF Multiple Range:        TUKEY Multiple Range:

    Indian ==                    Indian ==

       Med ┐                        Med ==
    Atl-ne │                     Atl-ne ┐
     Pac-w │                      Pac-w │
   Pac-cen │                    Pac-cen │
    Atl-nw │                     Atl-nw │
    Pac-ne ┘                     Pac-ne ┘

   Per-Red ==                   Per-Red ==
```

FIGURE 22. Results of stratification on location based on months. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

```
---------------------------------------
Stratification Results
---------------------------------------
Data Set:   MONTHS
Variable:   PAYGRADE
---------------------------------------

REGWF Multiple Range:      TUKEY Multiple Range:

  Enlisted ==                 Enlisted ==

   Officer ==                  Officer ==
```

FIGURE 23. Results of stratification on pay grade based on months. Strata are bracketed. Double line indicates the variable is the sole inhabitant of the stratum.

4.5 Time Series Analysis

Figure 24 is the periodogram for total DNBI rates. There are striking peaks at around 3.5 days and at exactly 7.0 days. These periods are remarkably consistent over all five years of data that was analyzed; they seem to be related as much to clinic operation as to DBNI incidence patterns.

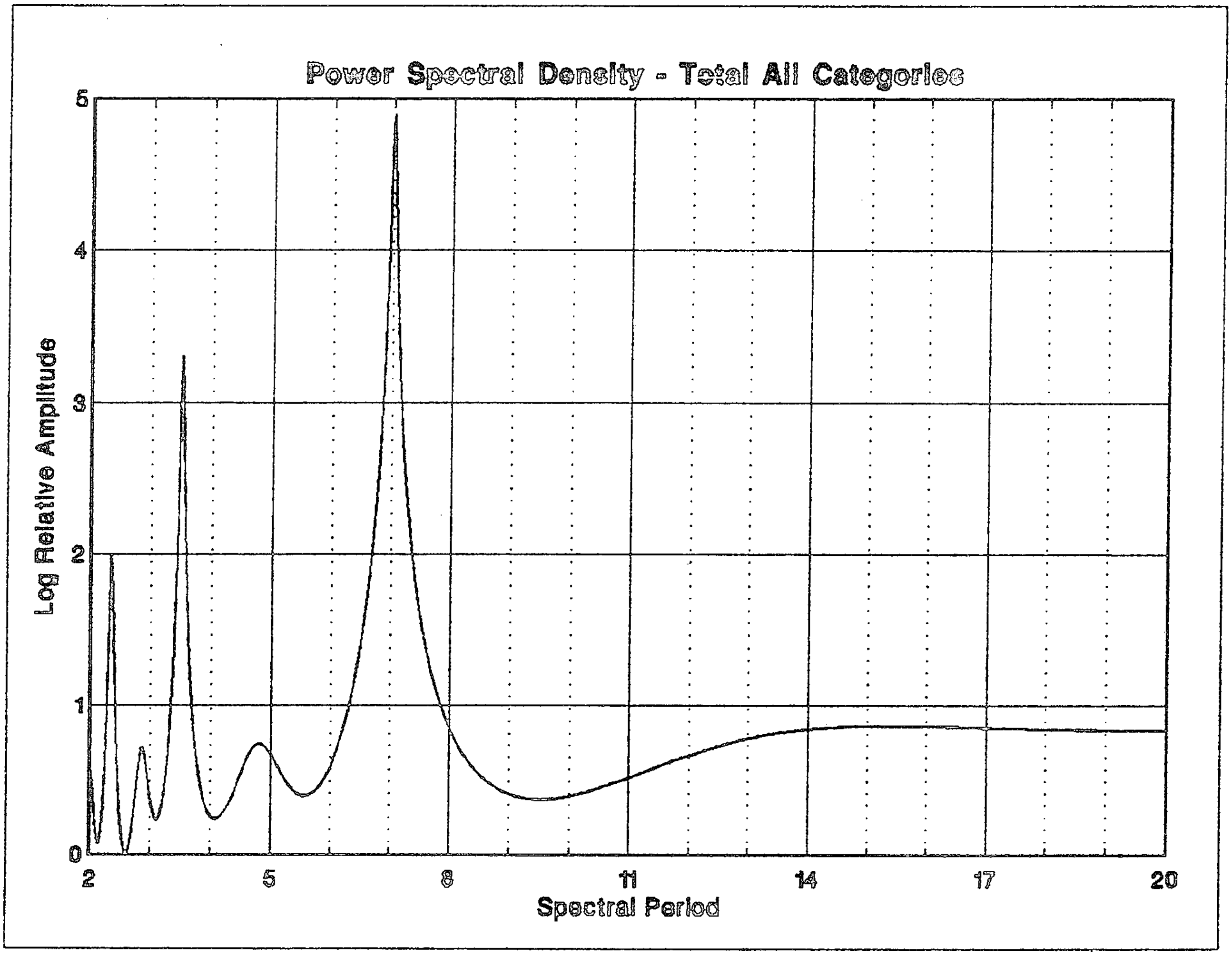

FIGURE 24. Spectrum for all ICD9-CM categories.

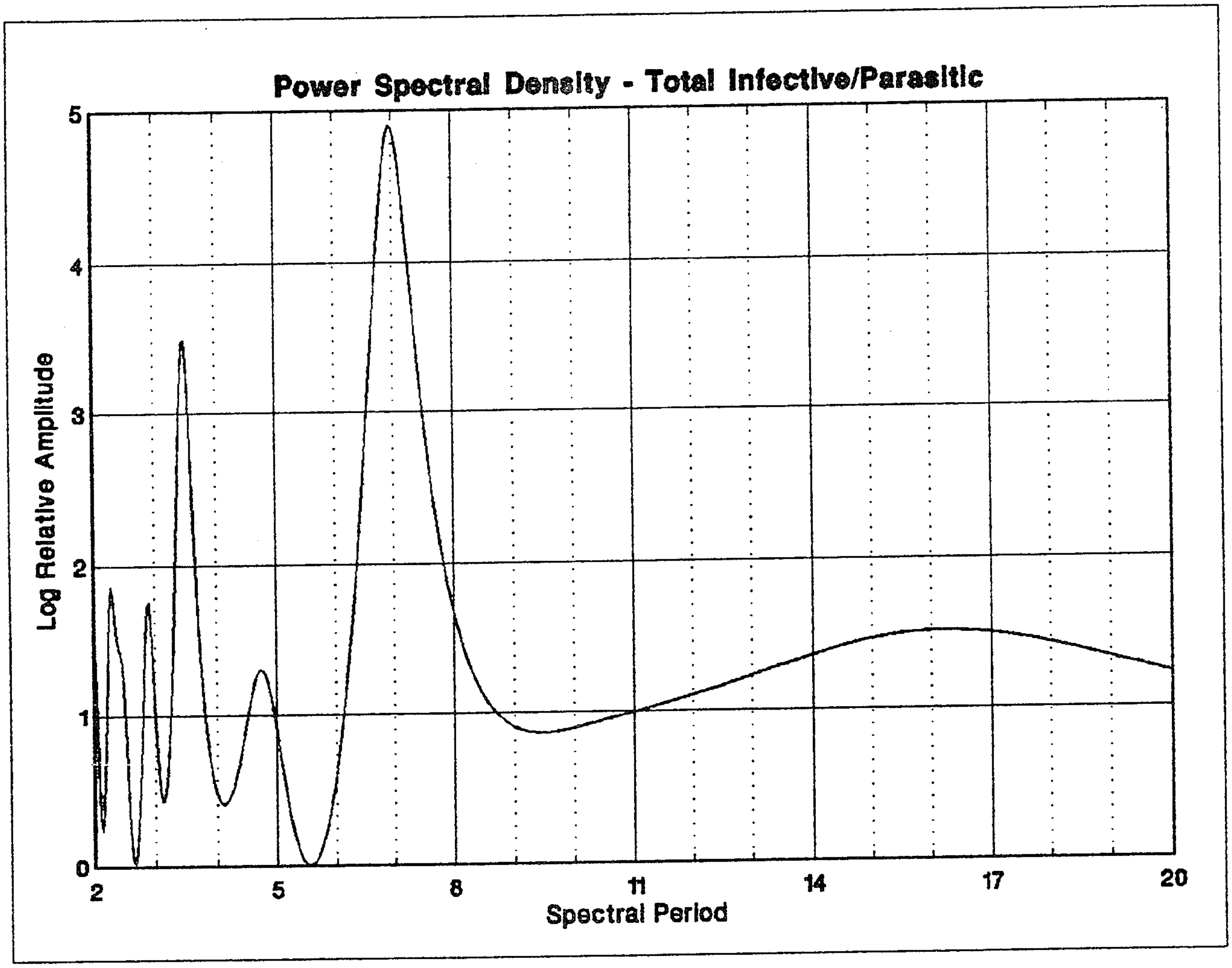

FIGURE 25. Spectrum for all infective/parasitic categories.

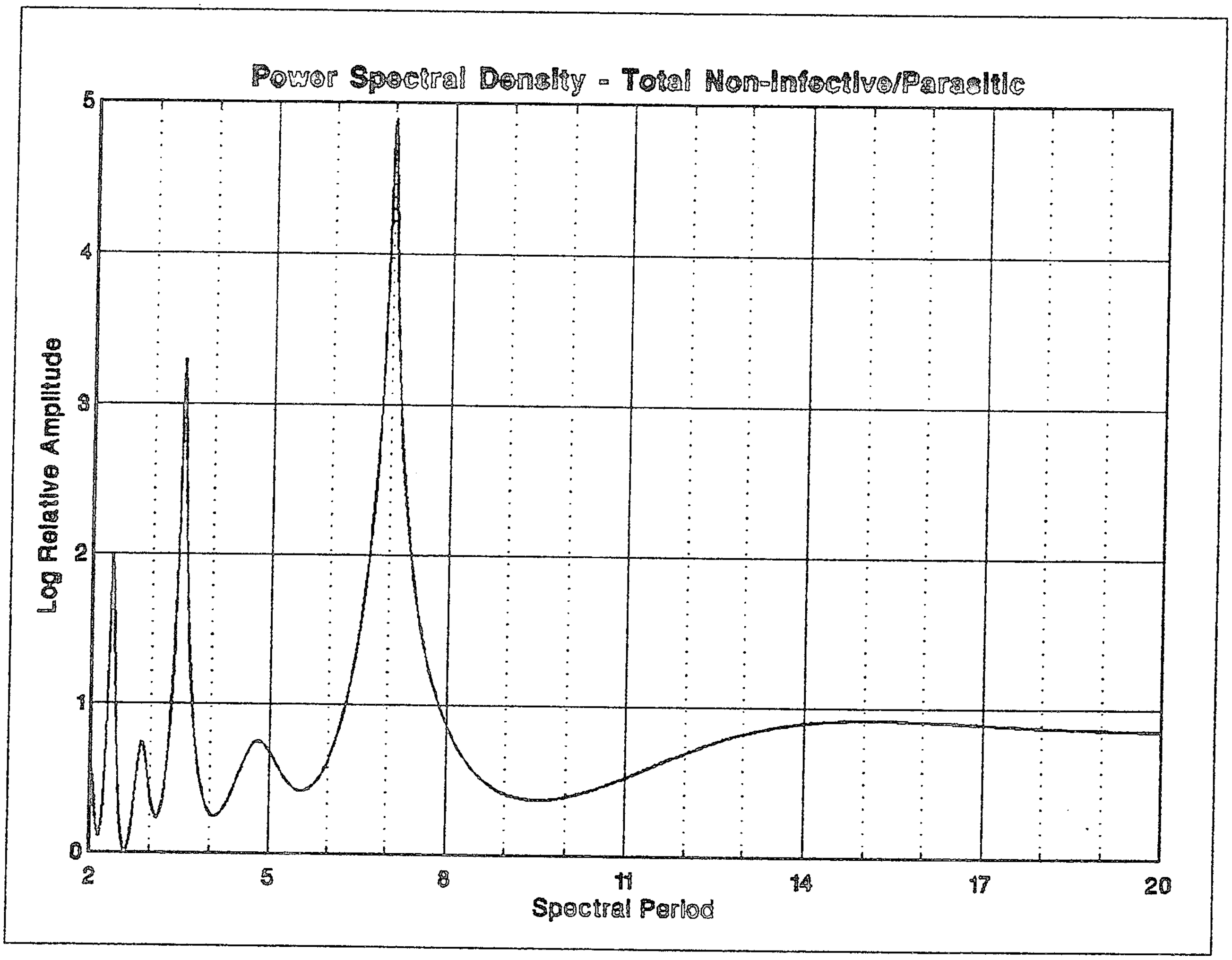

FIGURE 26. Spectrum for all except infective/parasitic categories.

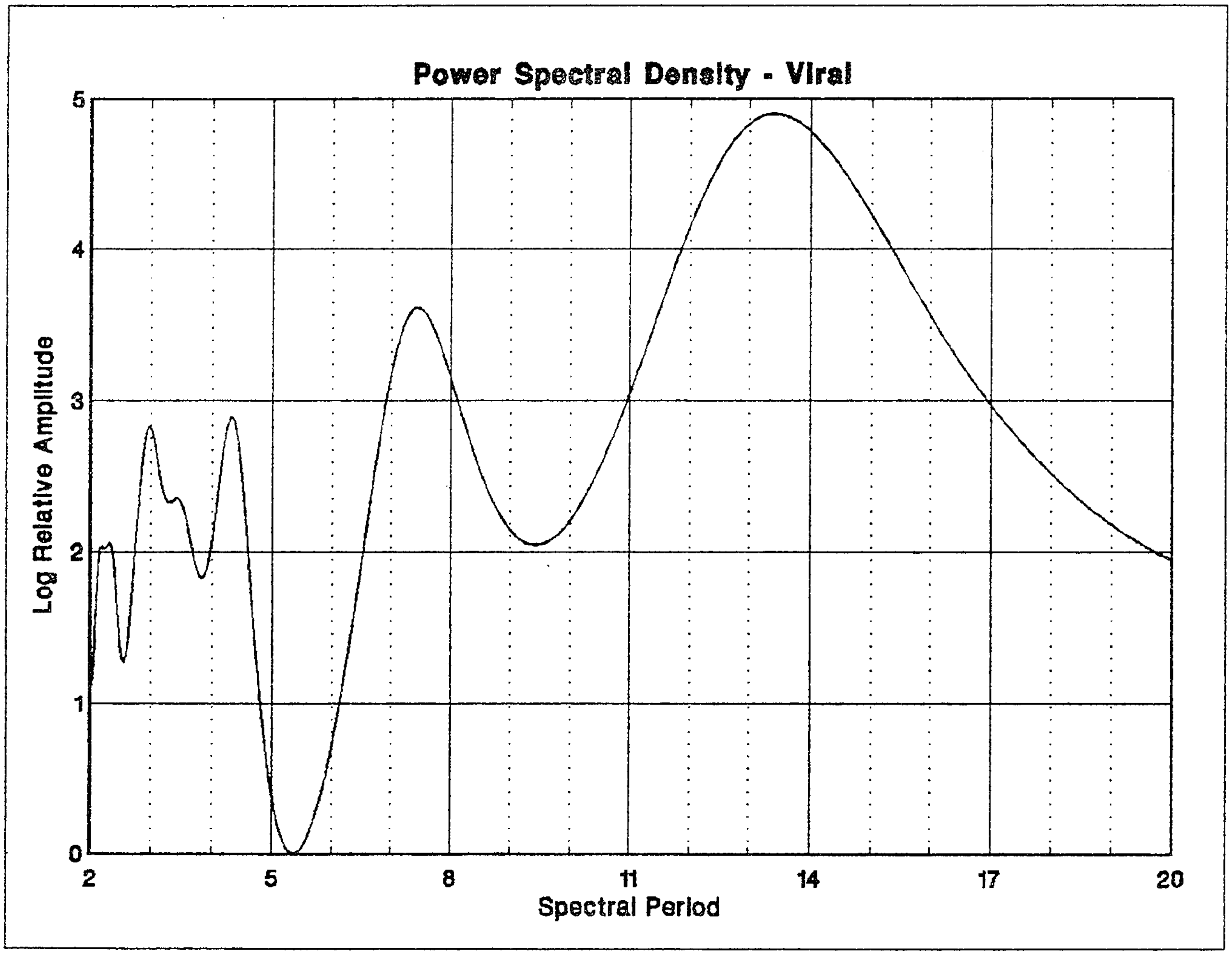

FIGURE 27. Spectrum for viral diseases.

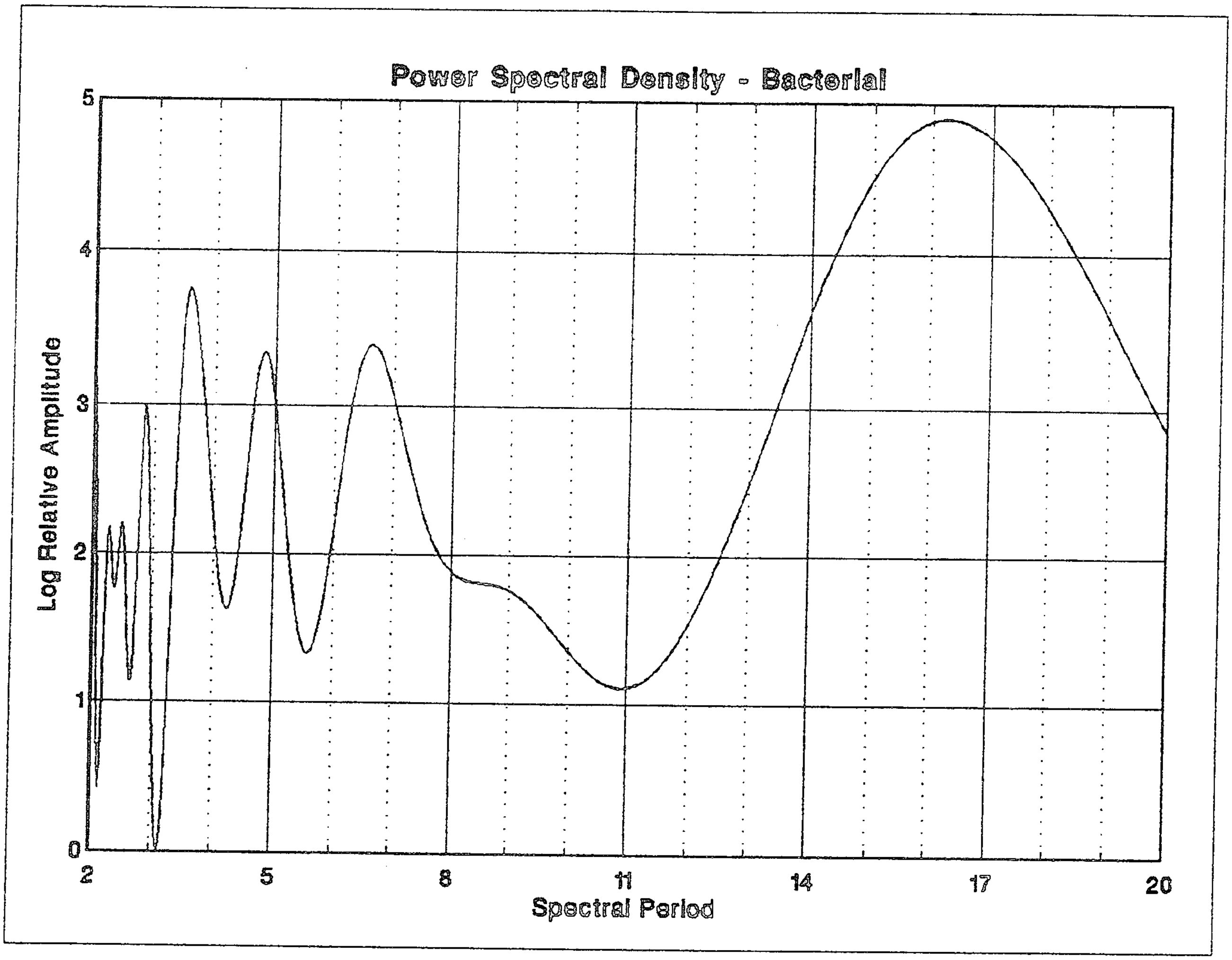

FIGURE 28. Spectrum for bacterial diseases.

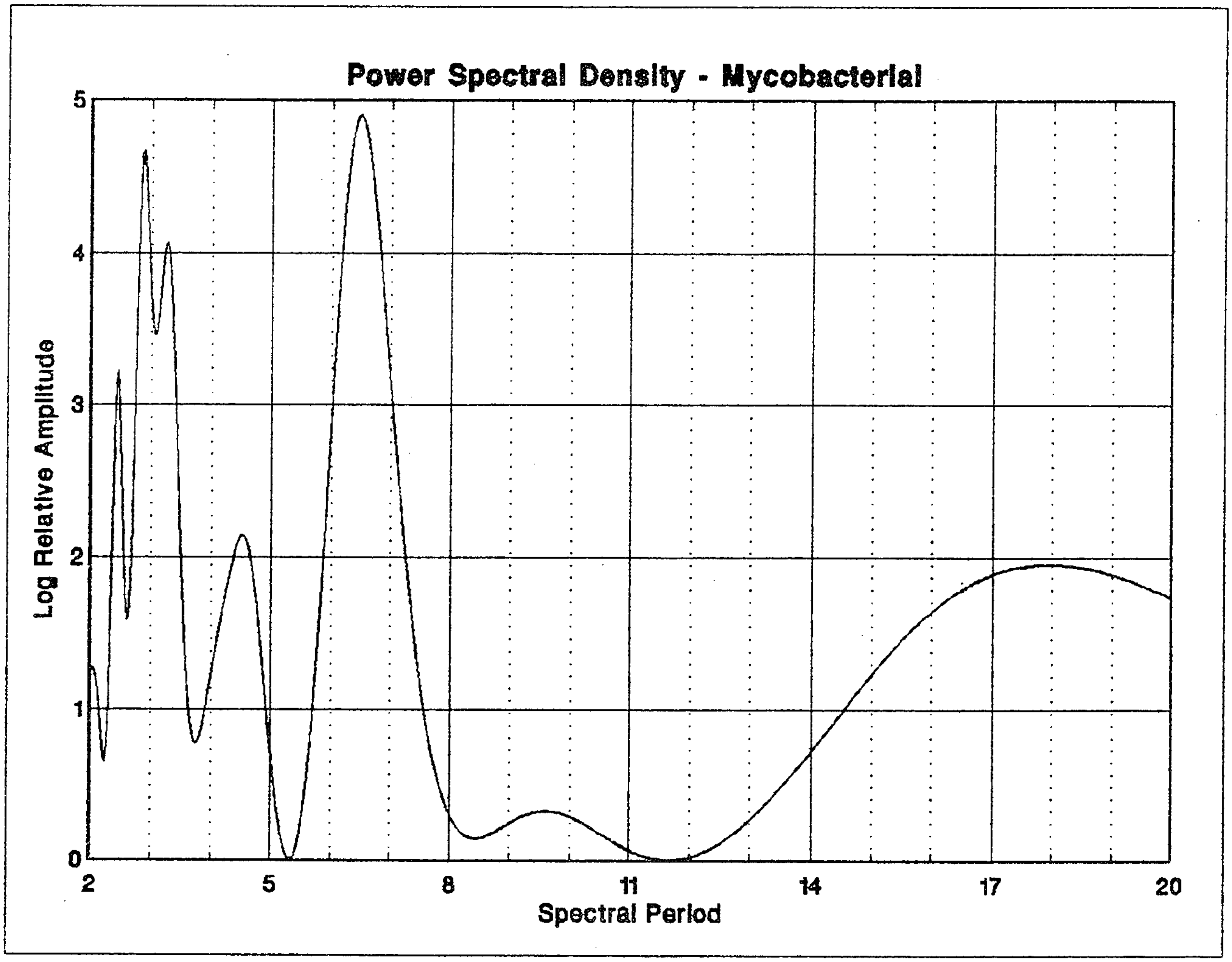

FIGURE 29. Spectrum for mycobacterial diseases.

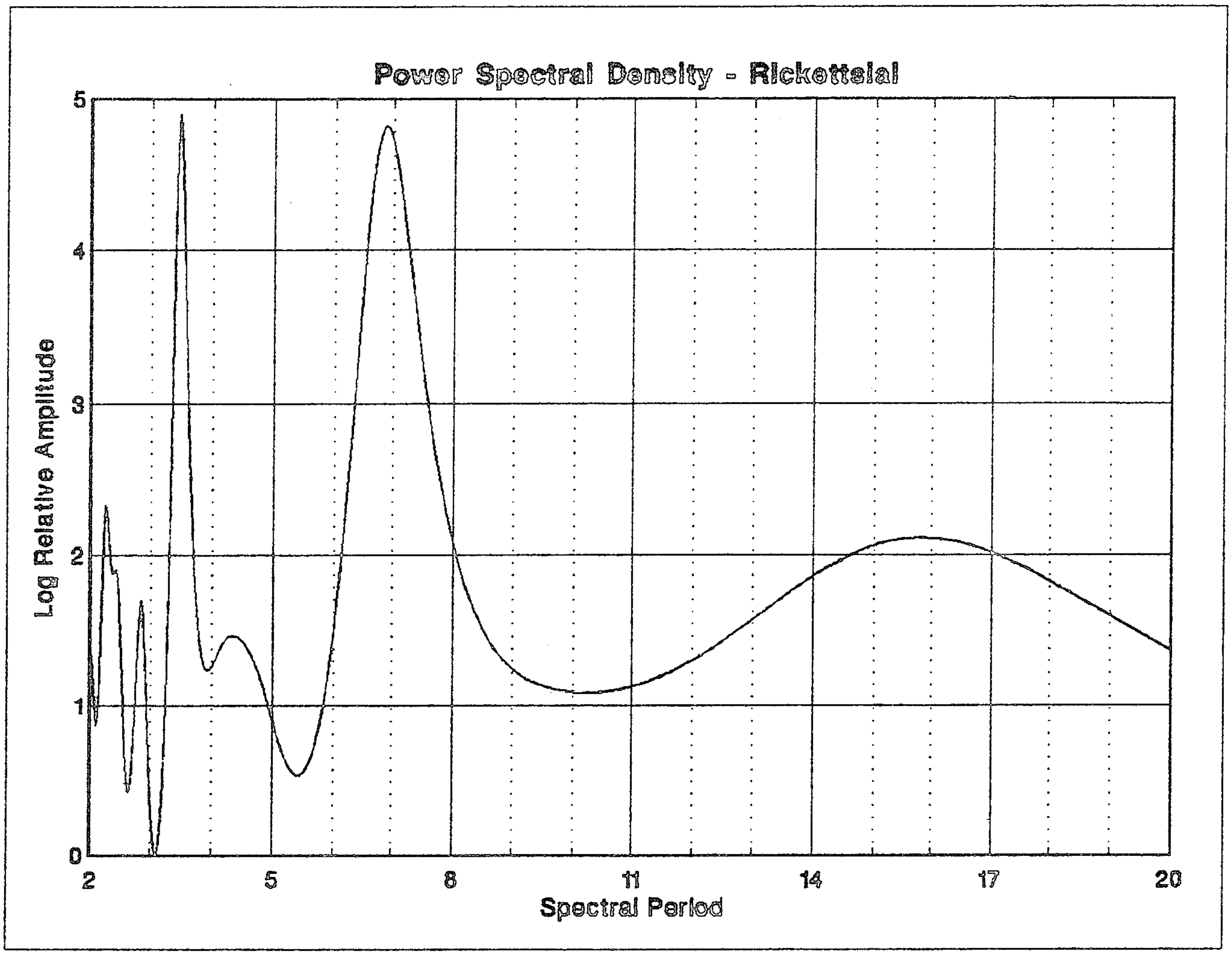

FIGURE 30. Spectrum for rickettsial diseases.

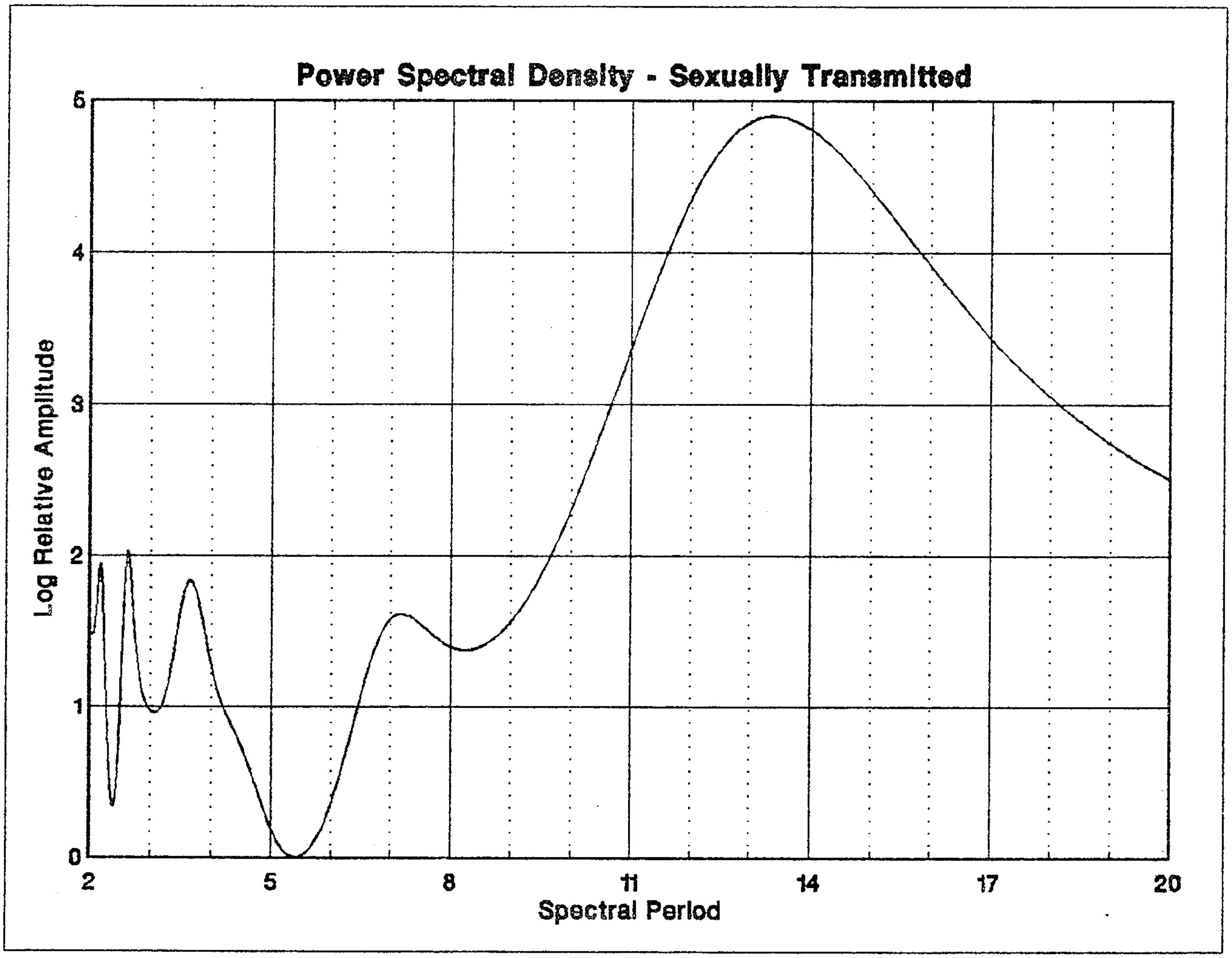

FIGURE 31. Spectrum for sexually transmitted diseases.

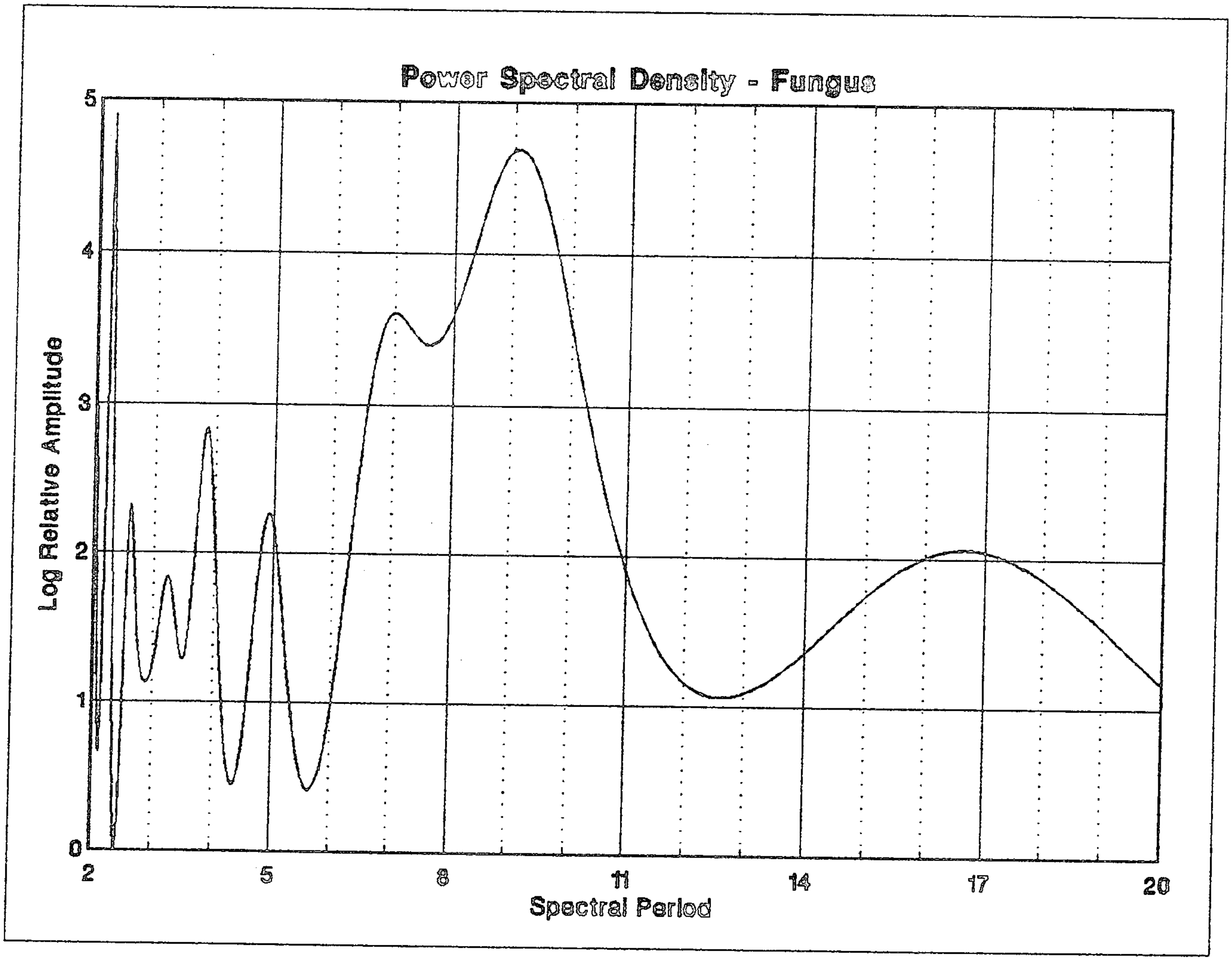

FIGURE 32. Spectrum for fungus diseases.

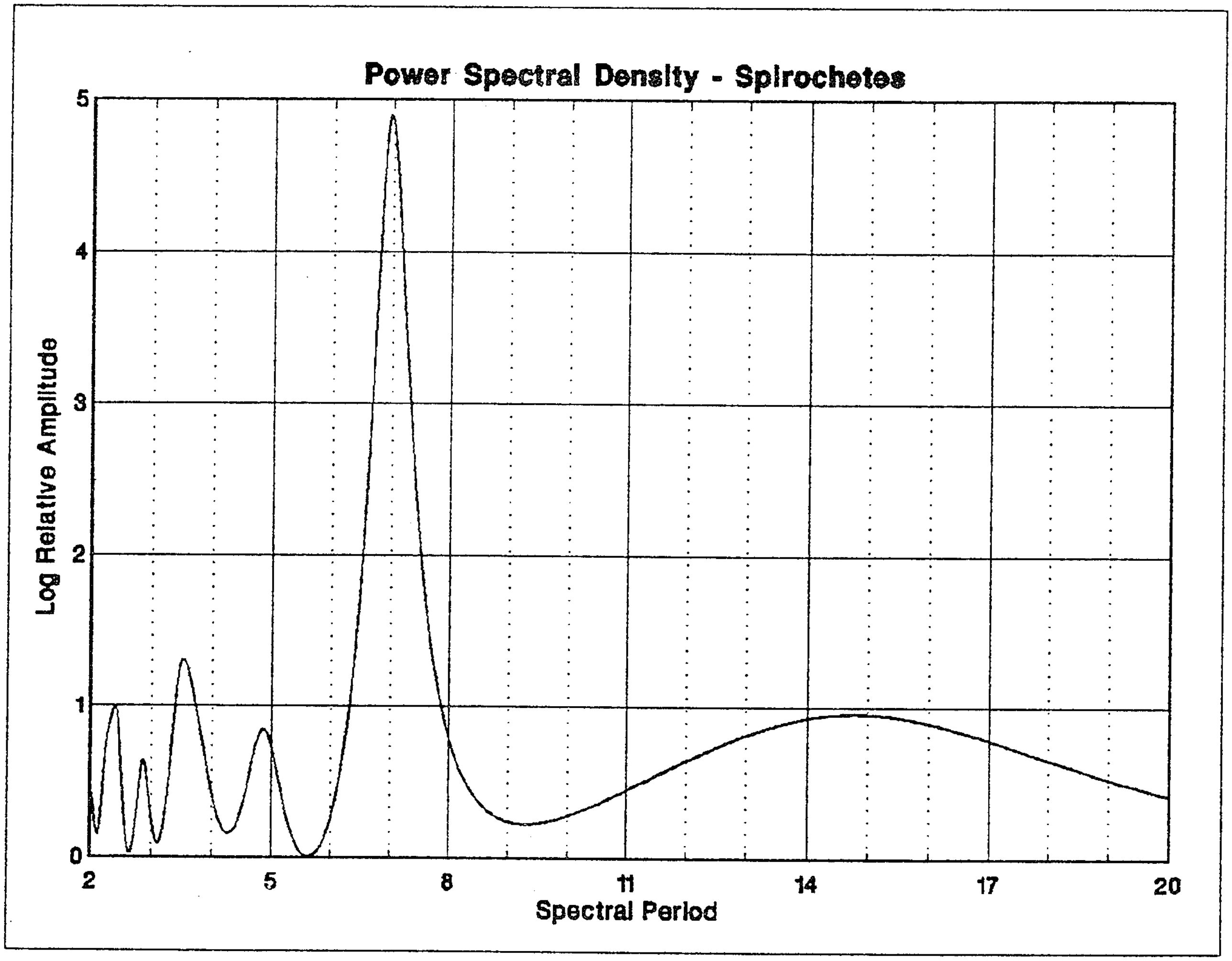

FIGURE 33. Spectrum for spirochete caused diseases.

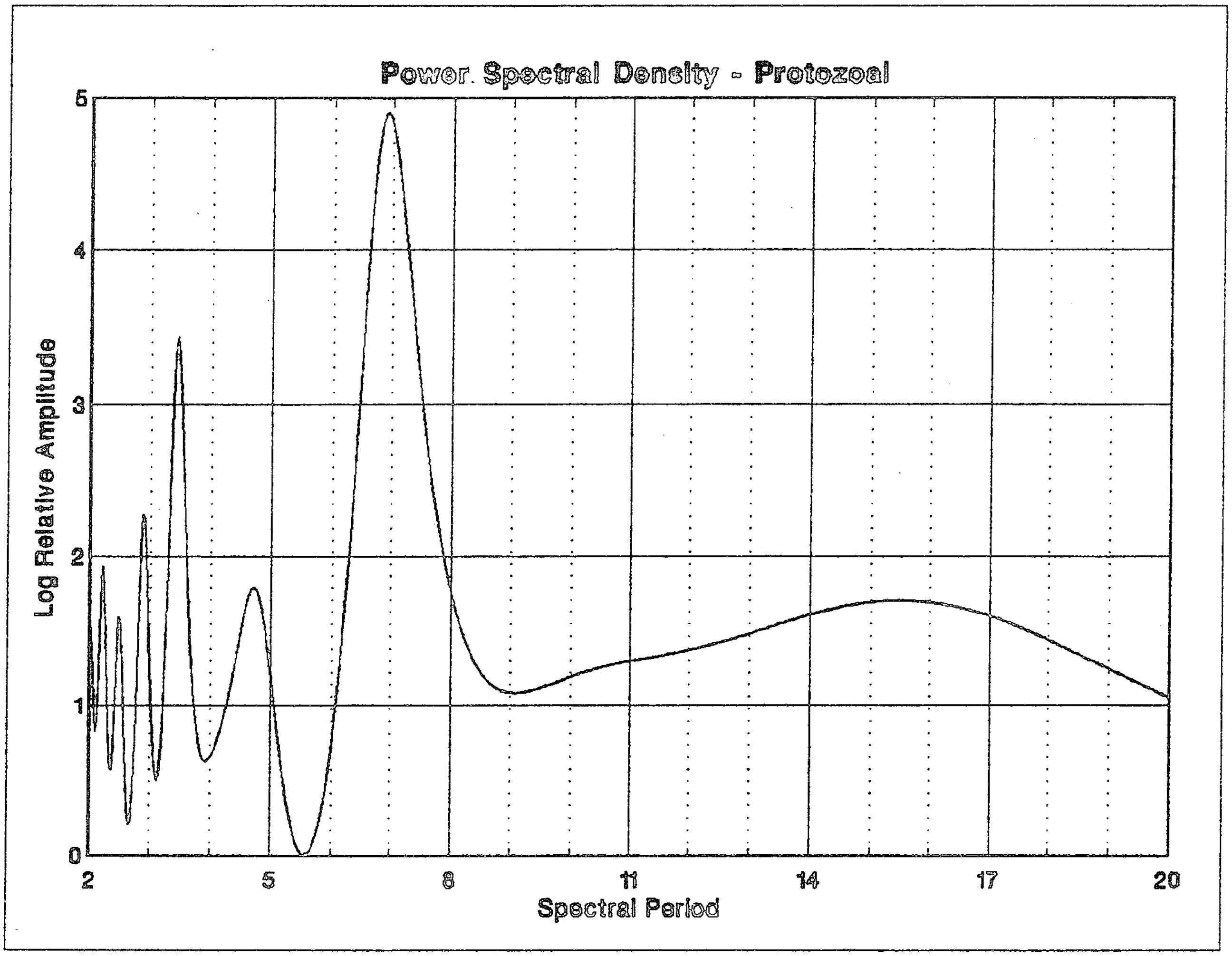

FIGURE 34. Spectrum for protozoal diseases.

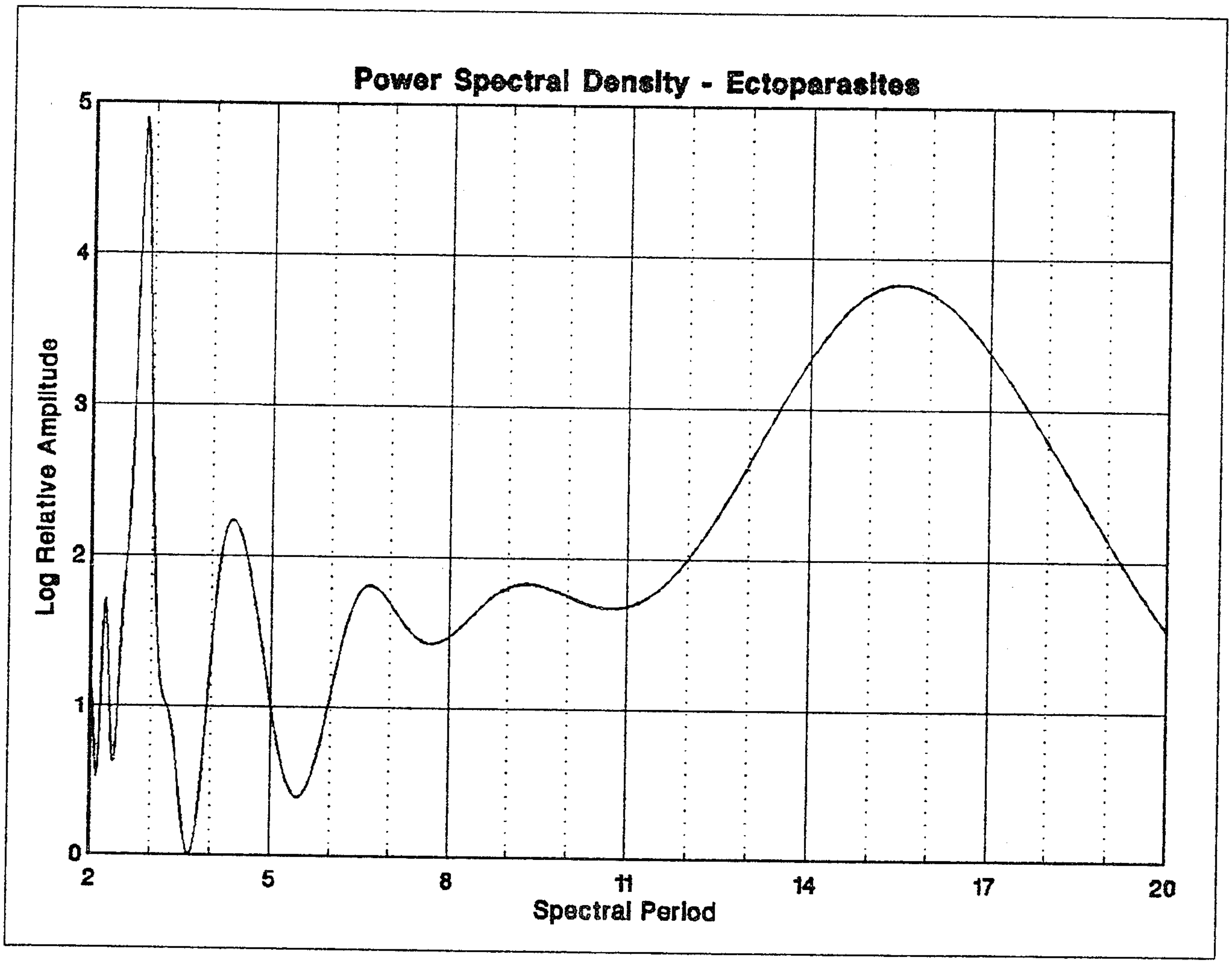

FIGURE 35. Spectrum for ectoparasites.

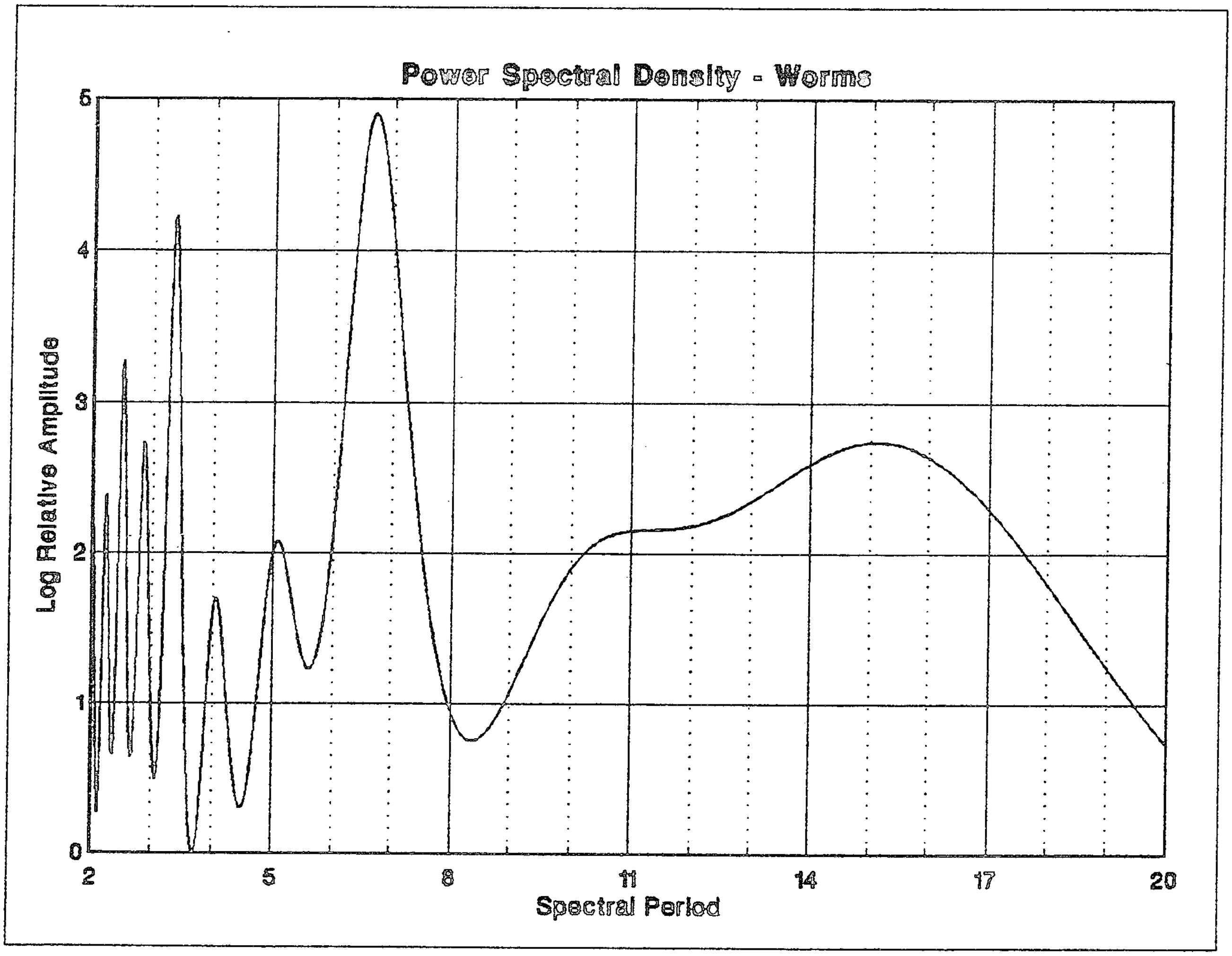

FIGURE 36. Spectrum for worm caused diseases.

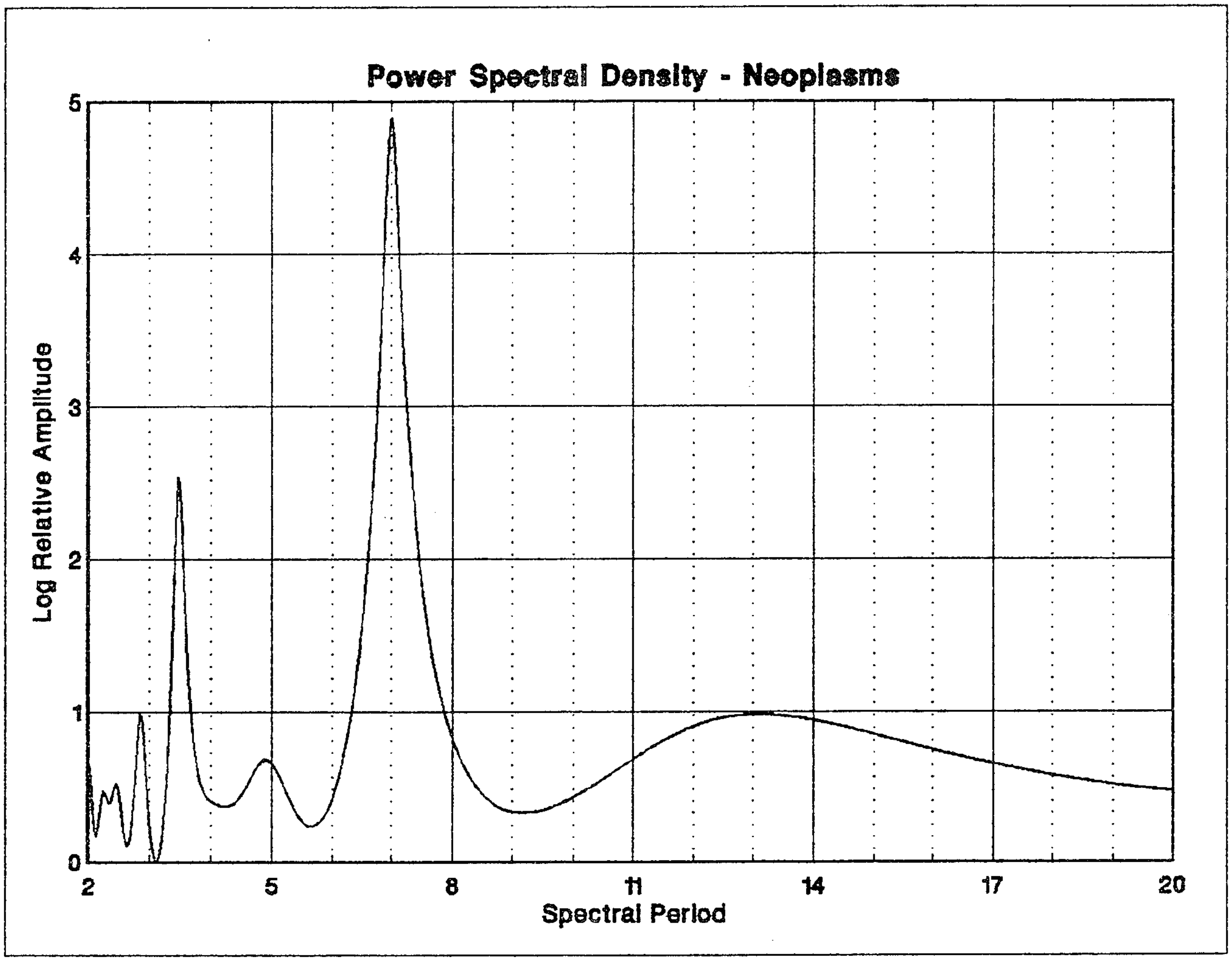

FIGURE 37. Spectrum for neoplams.

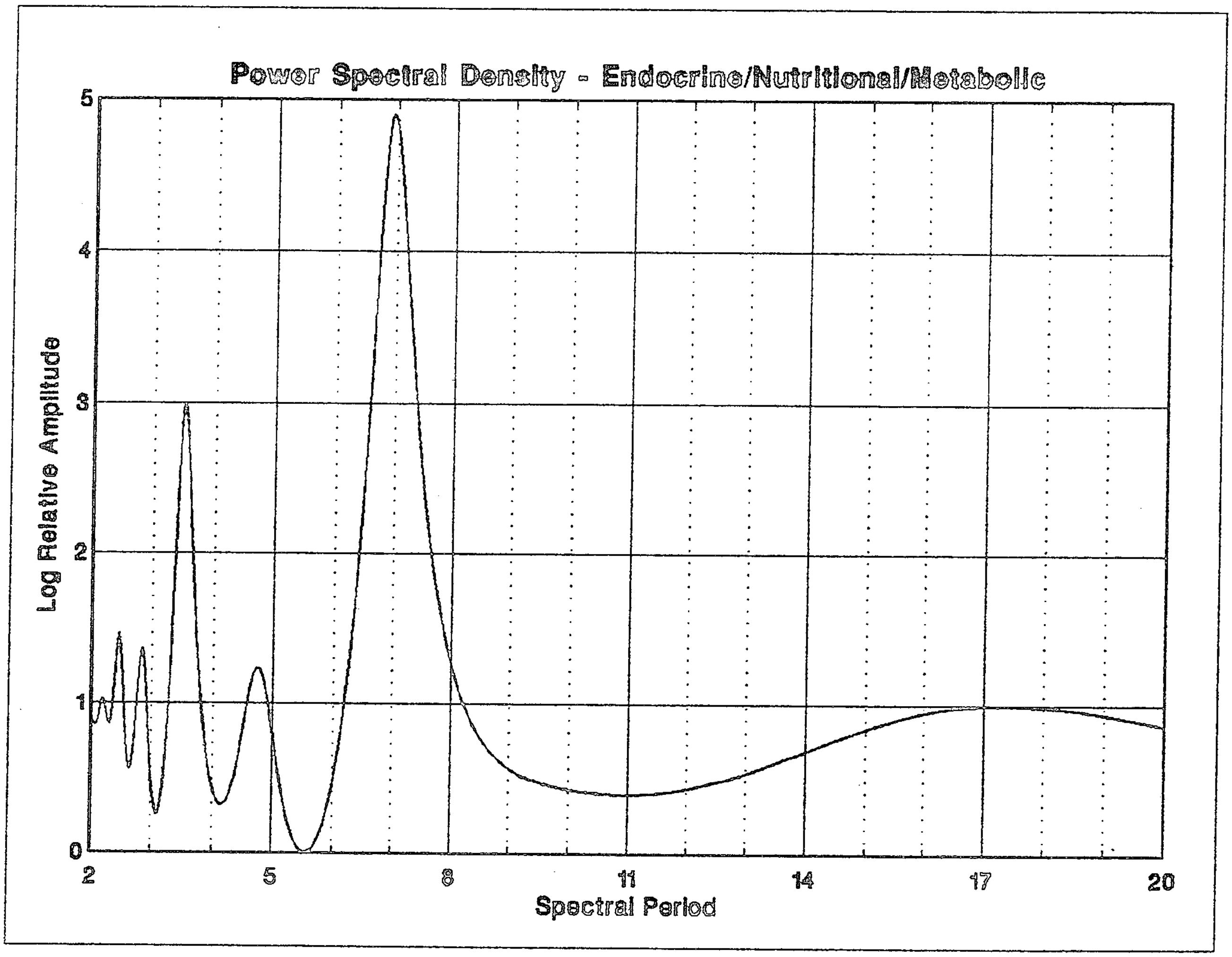

FIGURE 38. Spectrum for endocrine/nutritional/metabolic disorders.

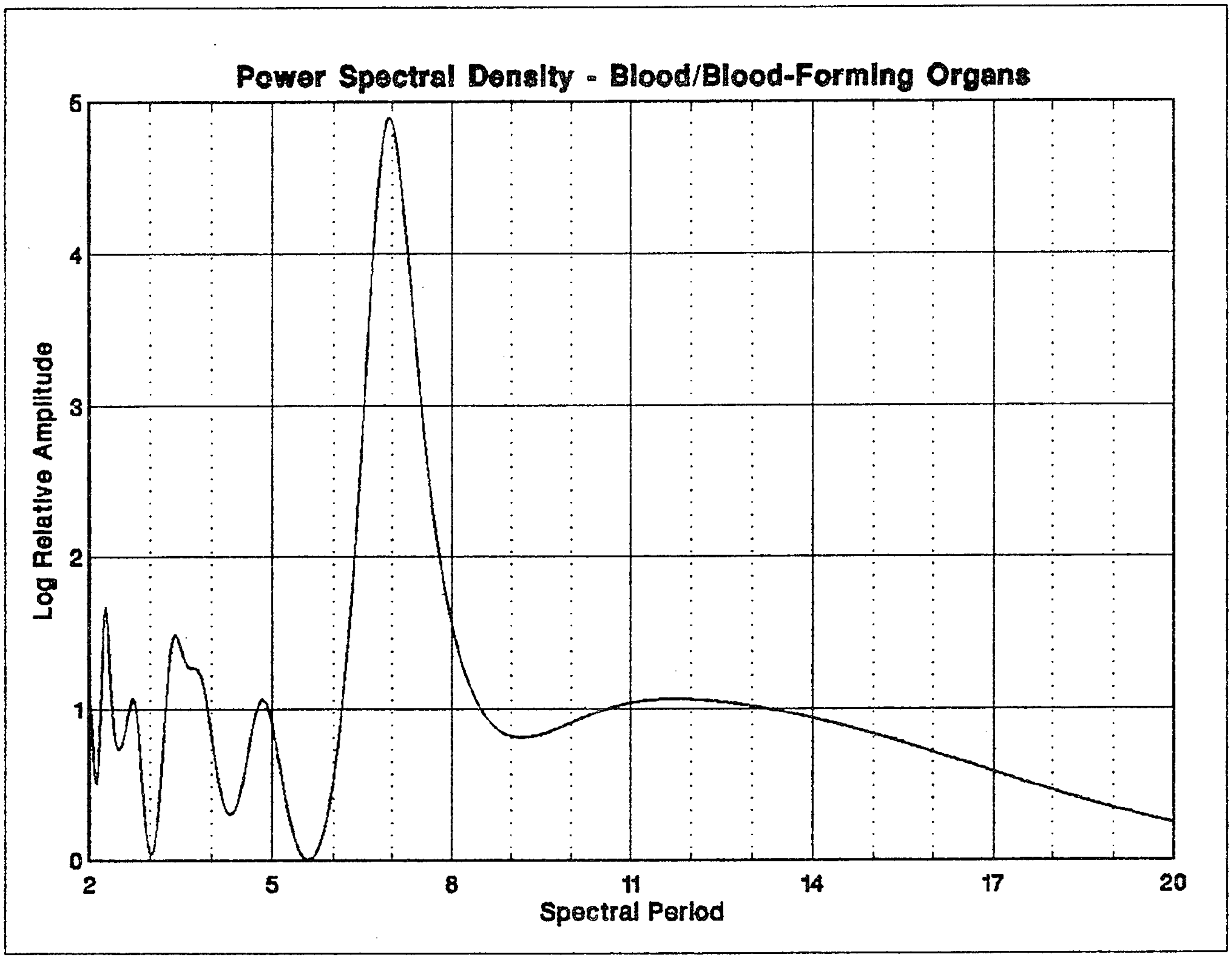

FIGURE 39. Spectrum for blood/blood-forming organ disorders.

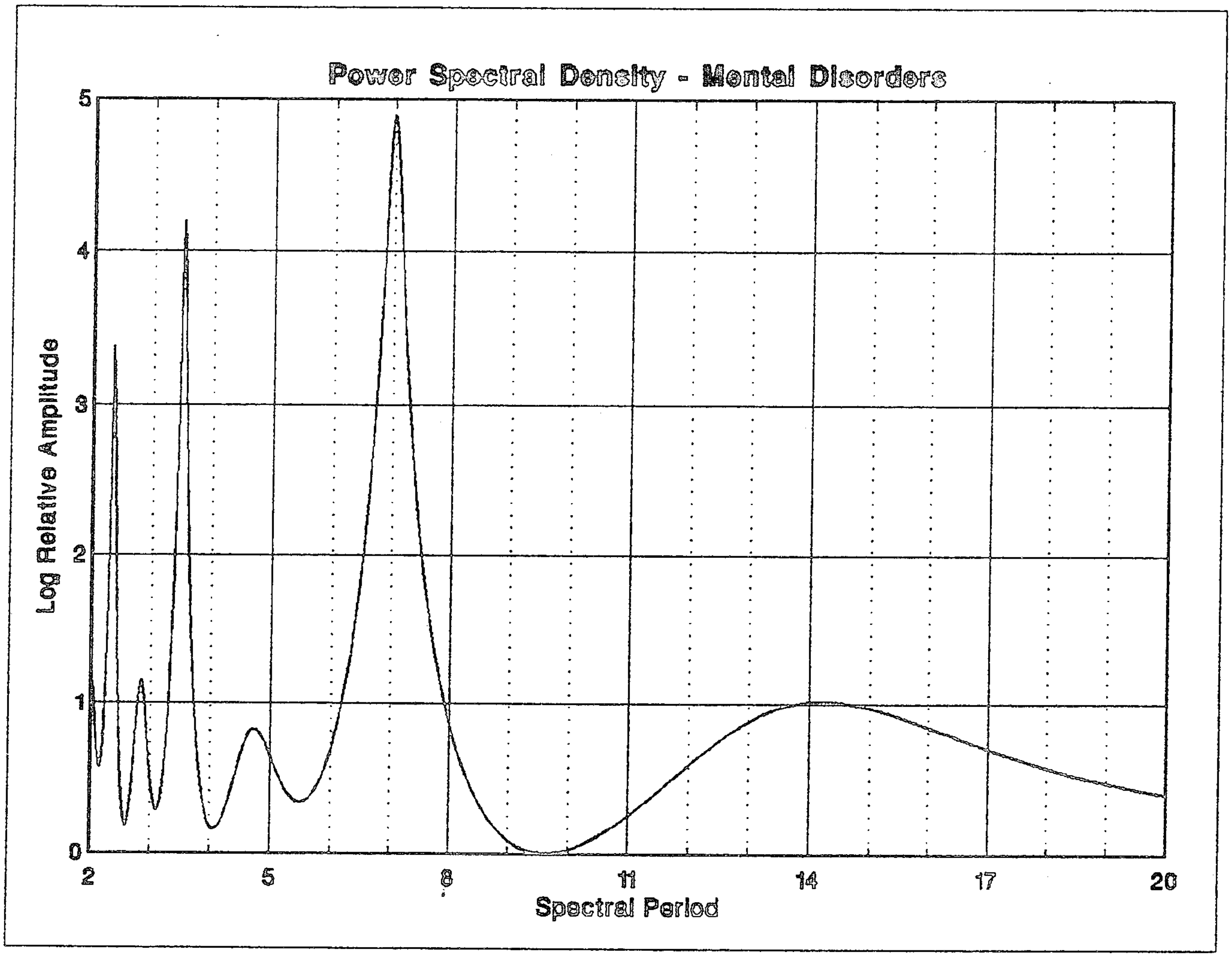

FIGURE 40. Spectrum for mental disorders.

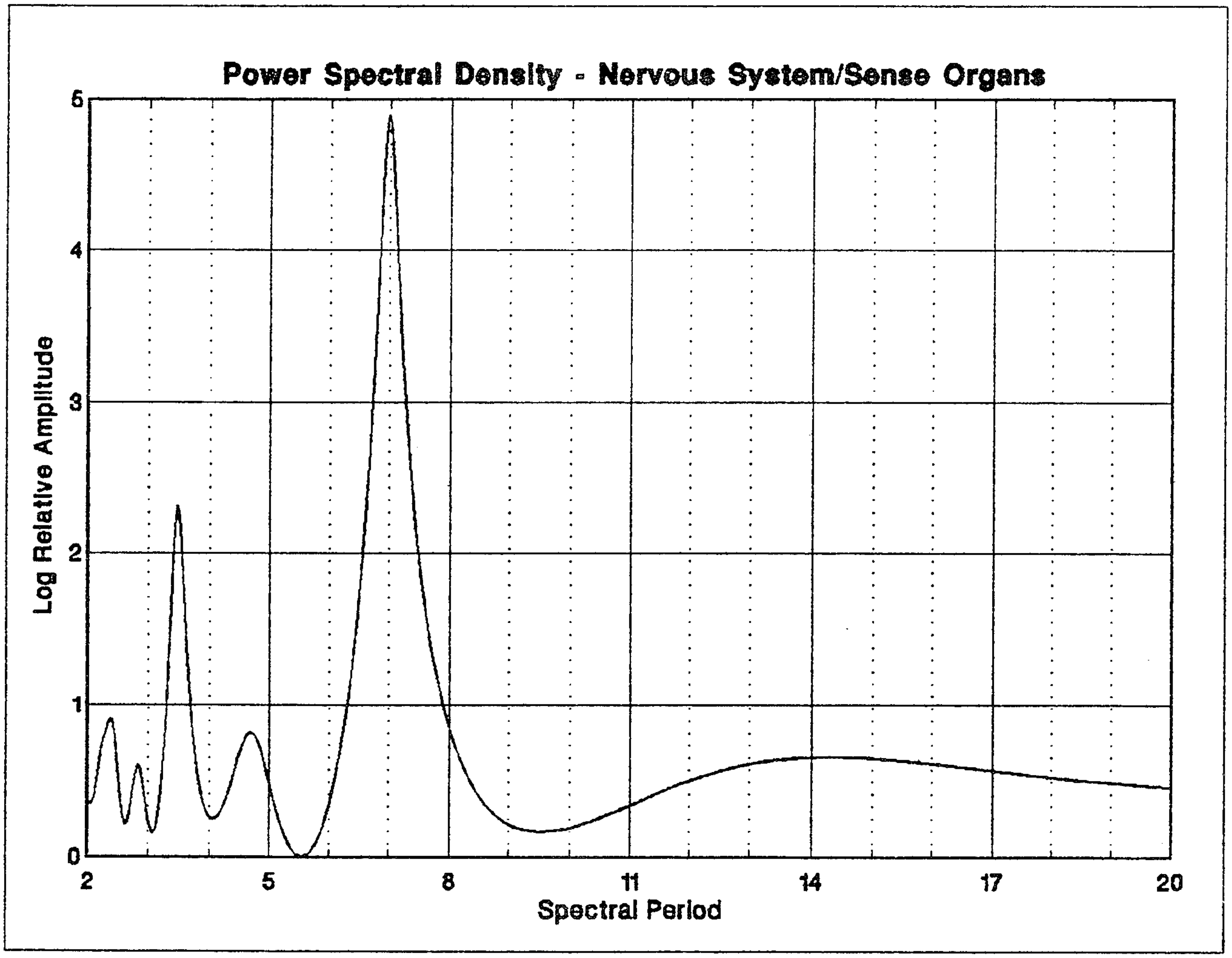

FIGURE 41. Spectrum for nervous system/sense organ disorders.

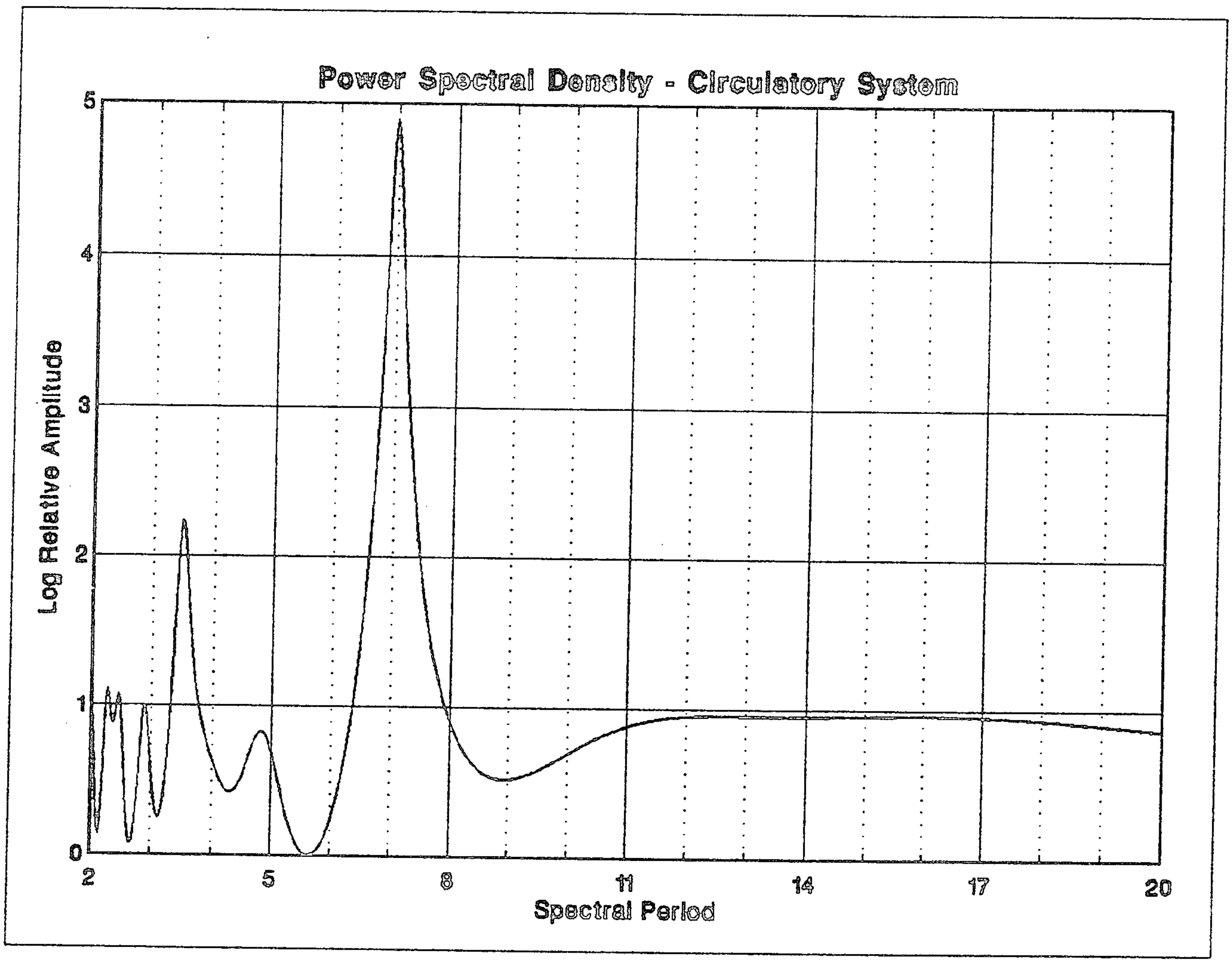

FIGURE 42. Spectrum for circulatory system disroders.

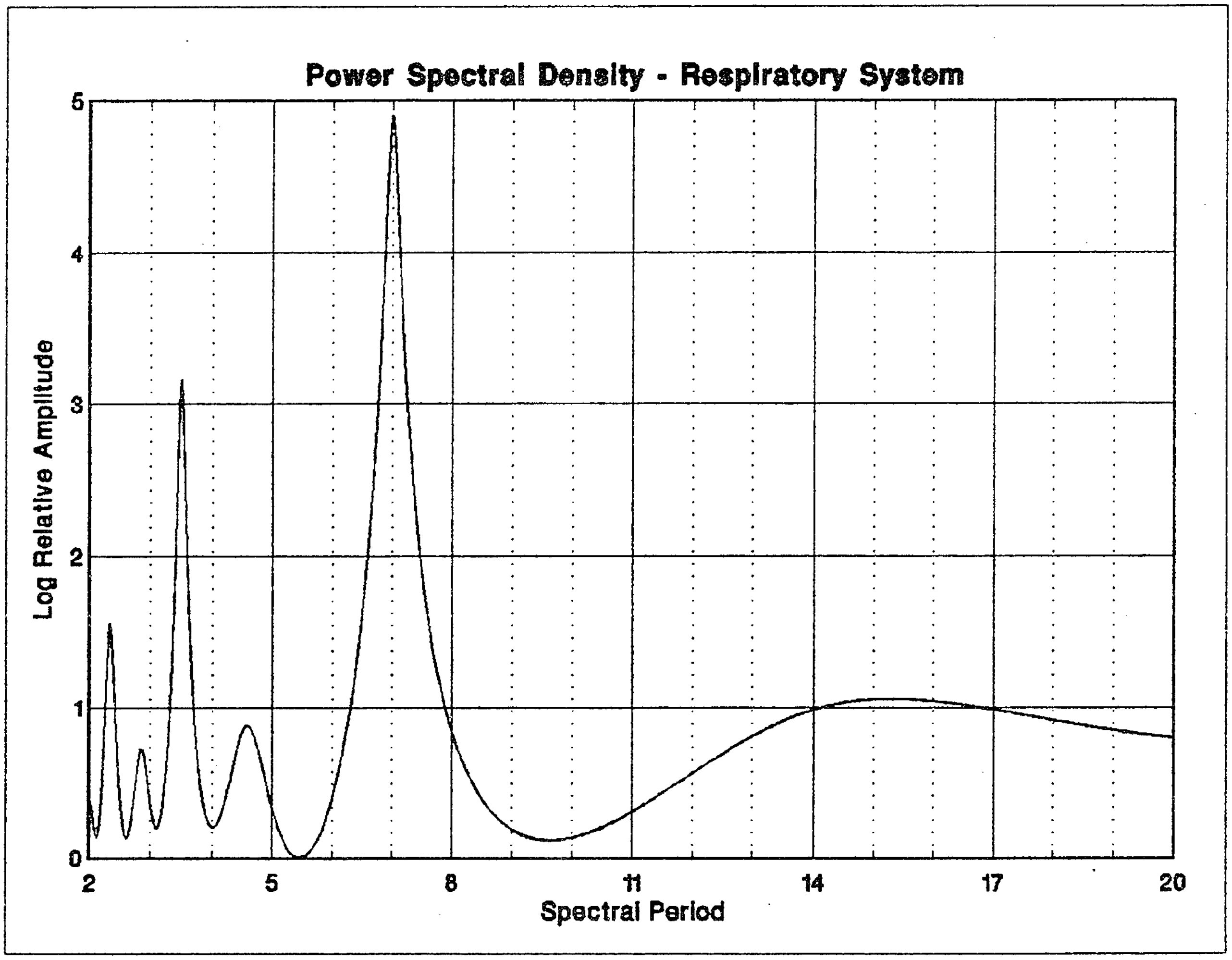

FIGURE 43. Spectrum for respiratory system disorders.

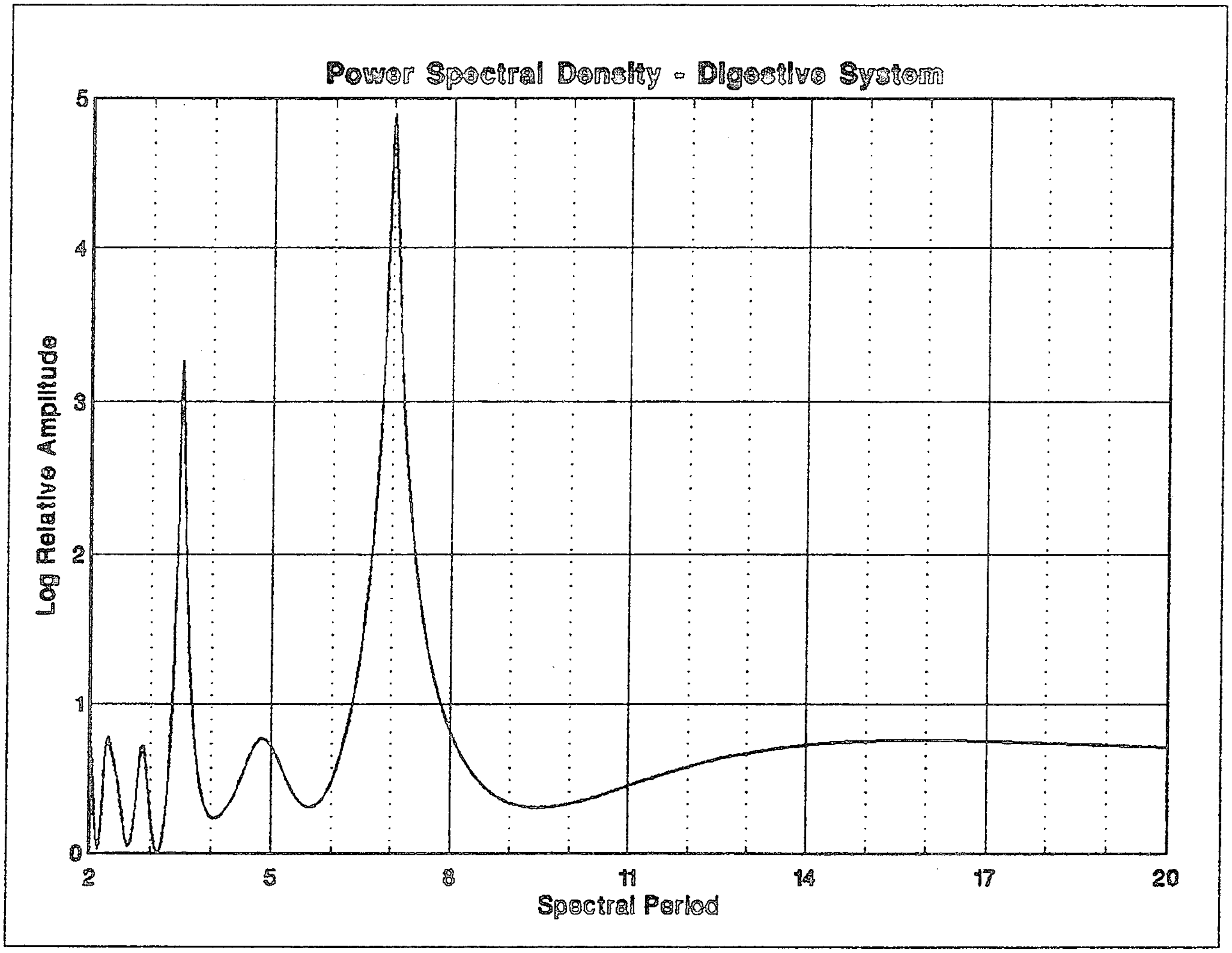

FIGURE 44. Spectrum for digestive system disorders.

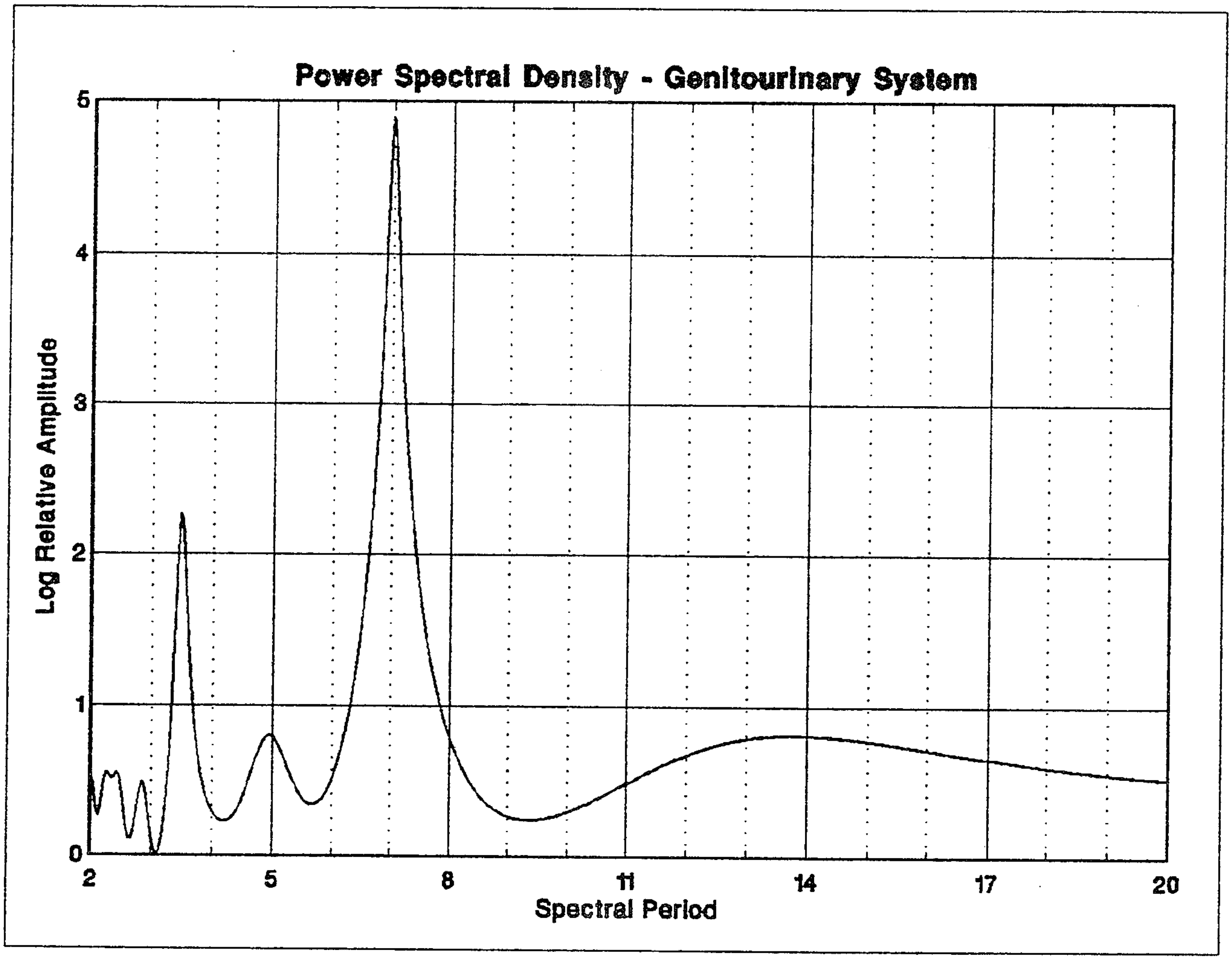

FIGURE 45. Spectrum for genitourinary system disorders.

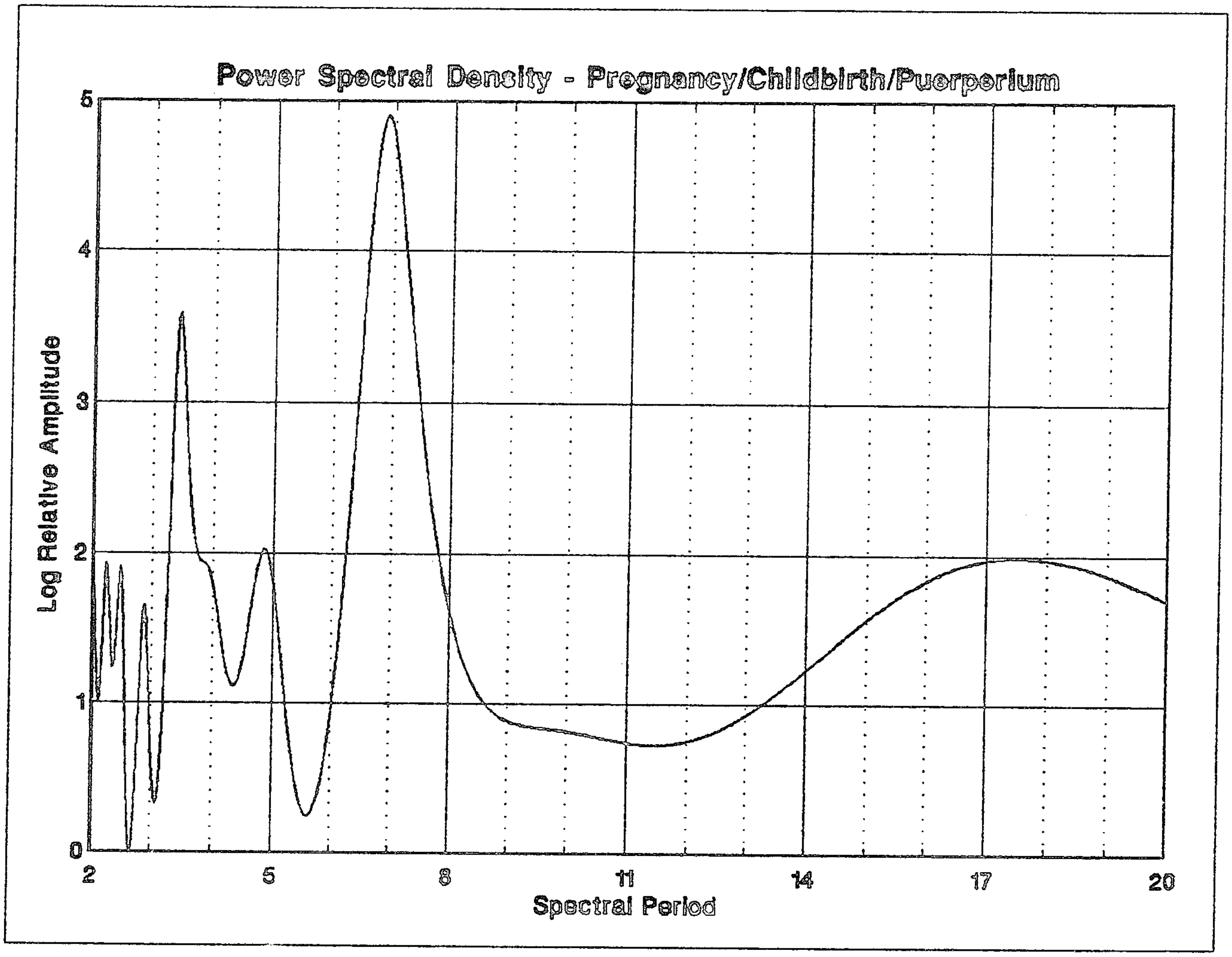

FIGURE 46. Spectrum for pregnancy/childbirth complications.

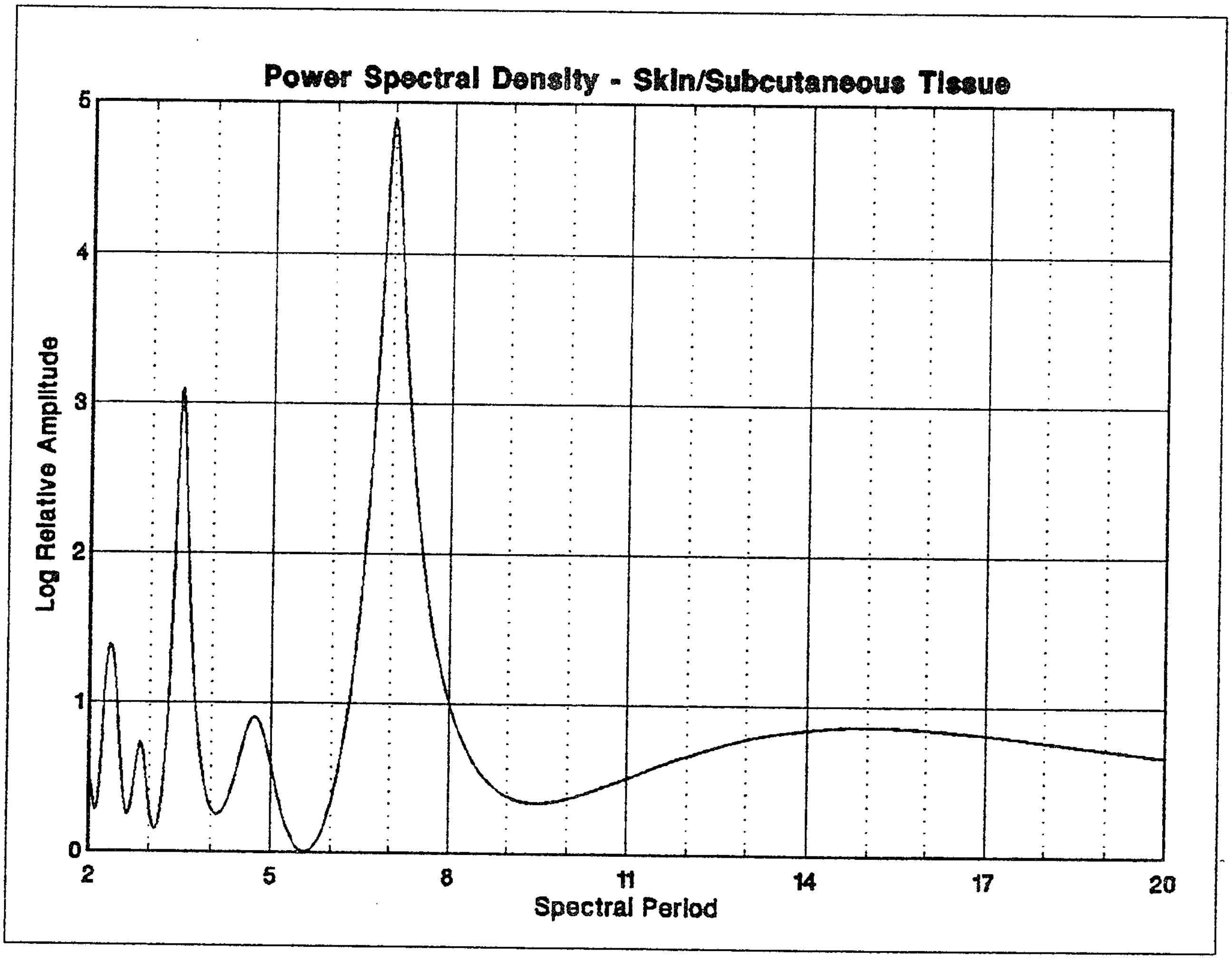

FIGURE 47. Spectrum for skin/subcutaneous tissue disorders.

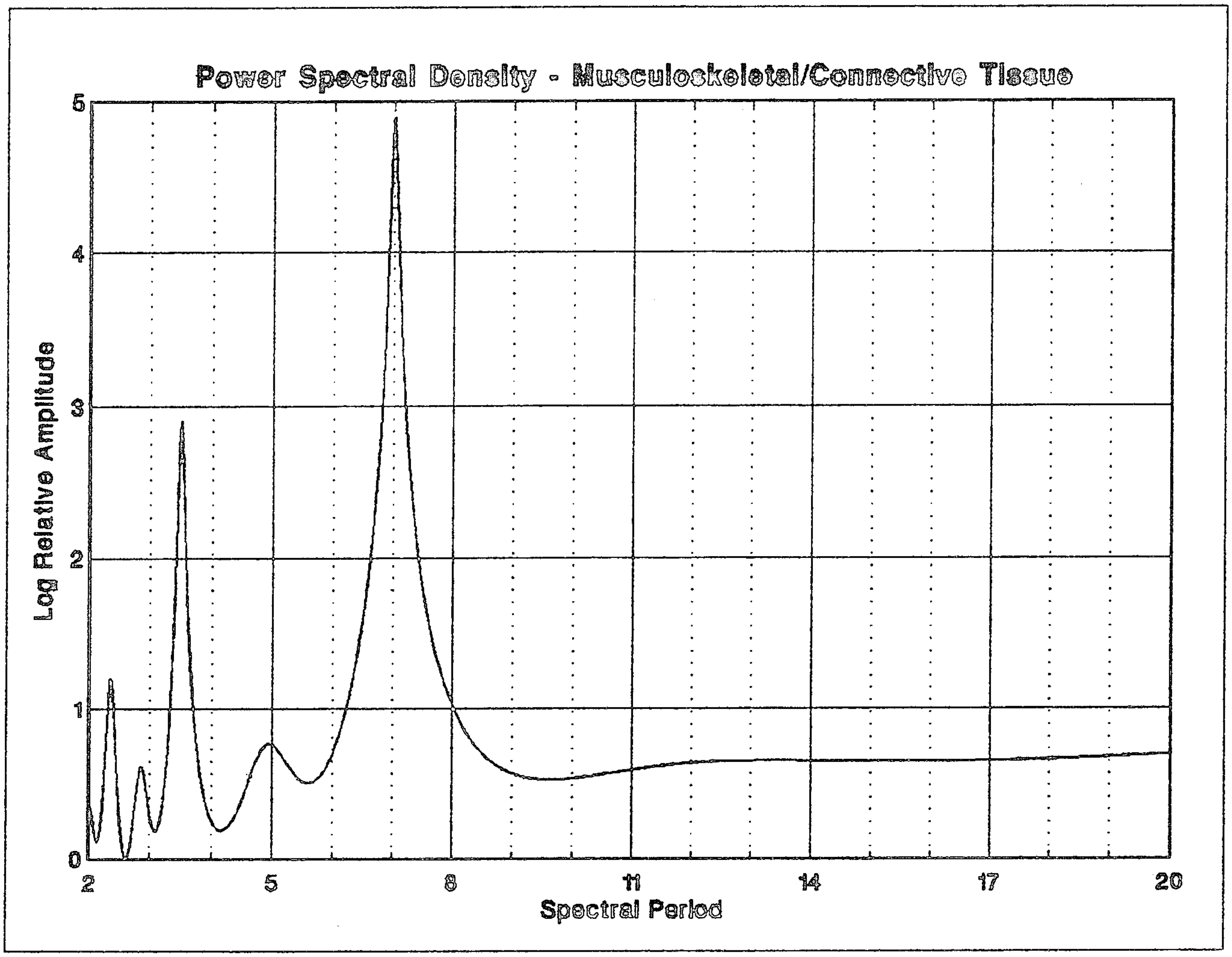

FIGURE 48. Spectrum for musculoskeletal/connective tissue disorders.

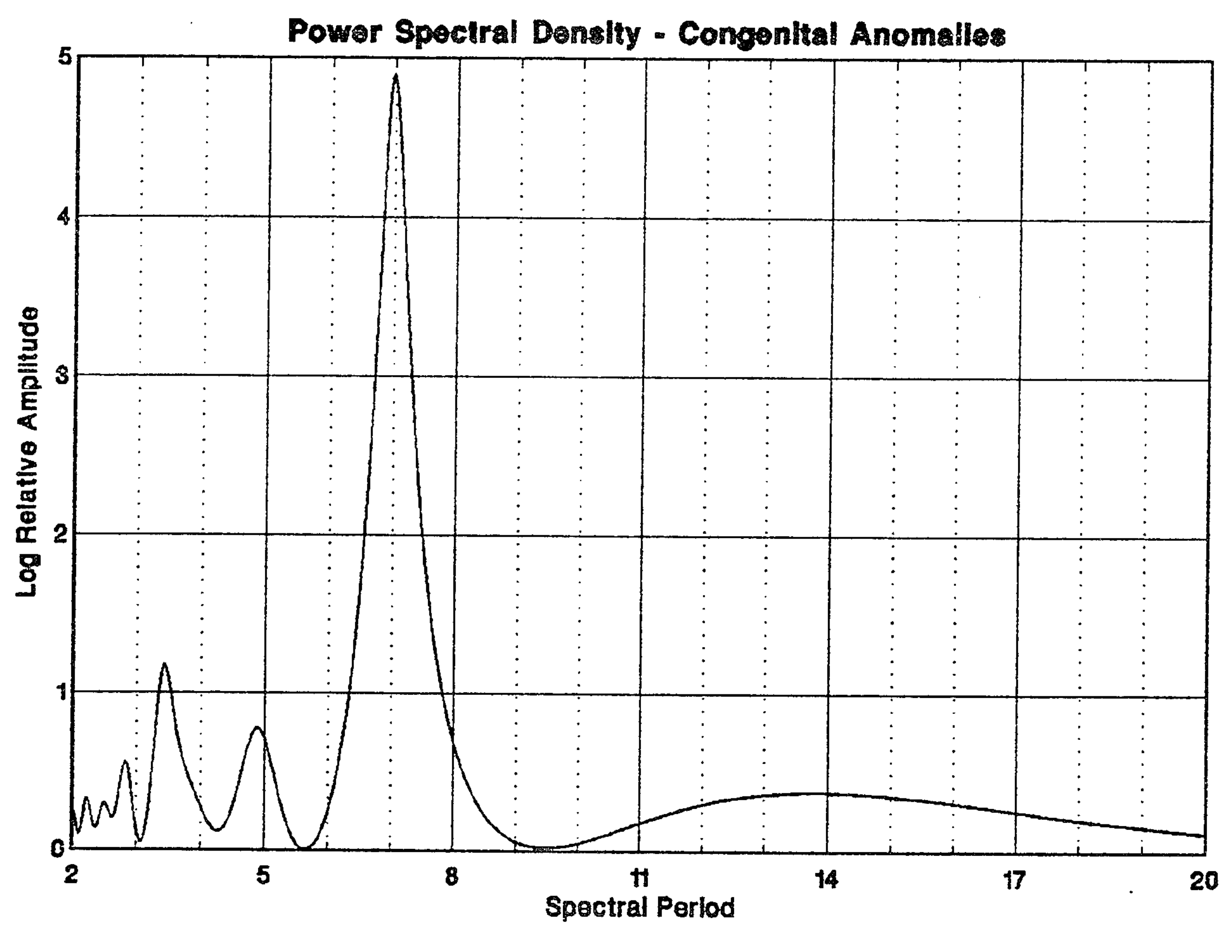

FIGURE 49. Spectrum for congenital anomalies.

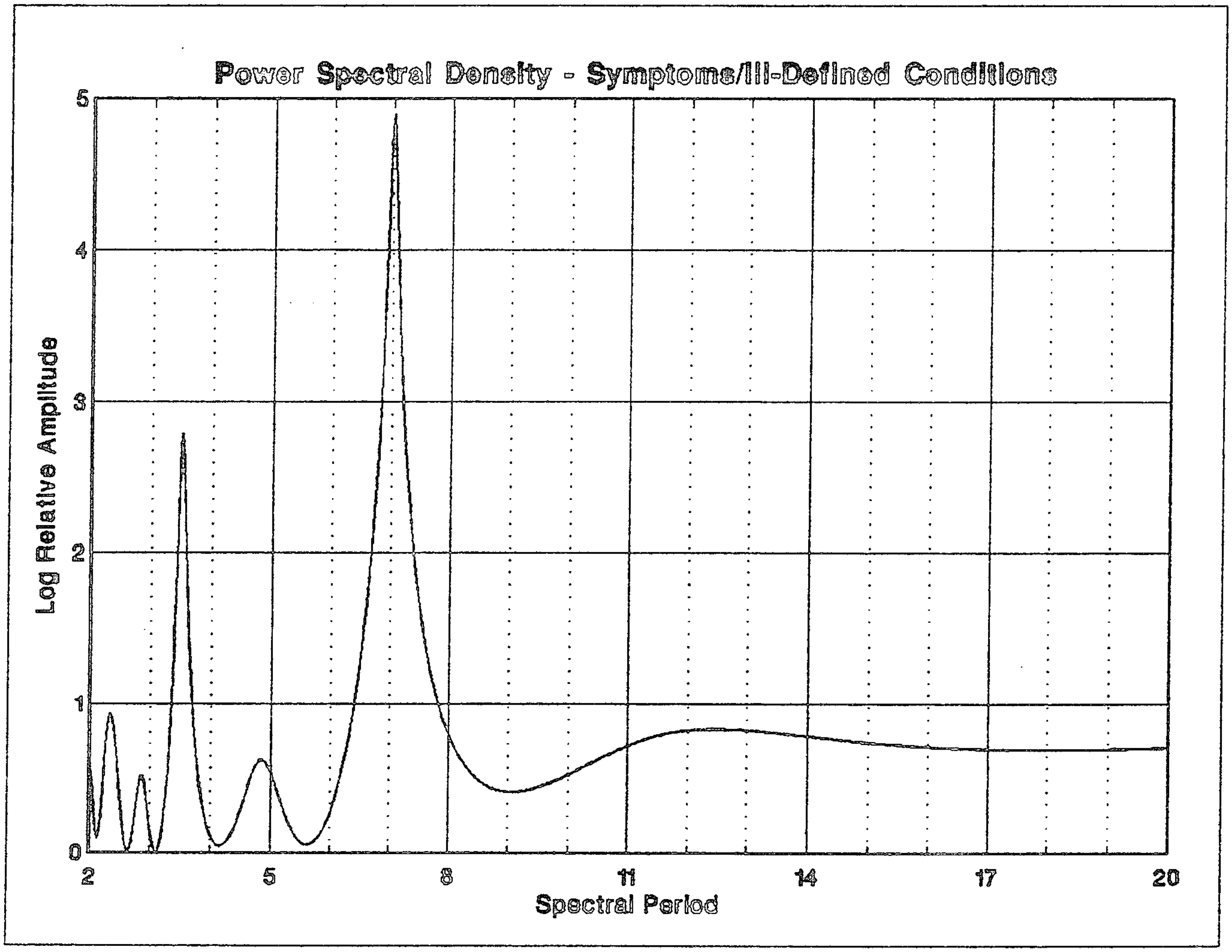

FIGURE 50. Spectrum for symptoms/ill-defined conditions.

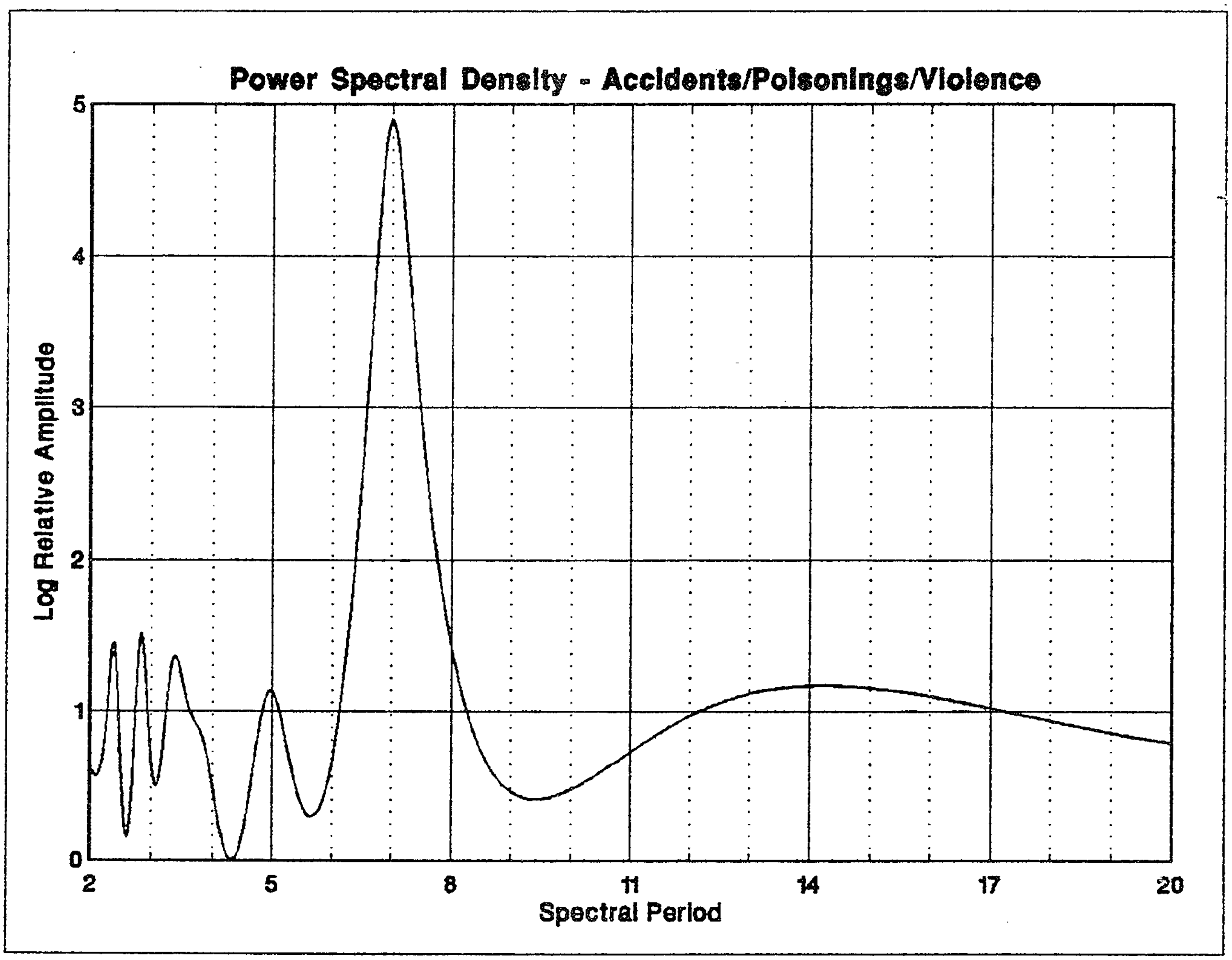

FIGURE 51. Spectrum for accidents/poisonings/violence.

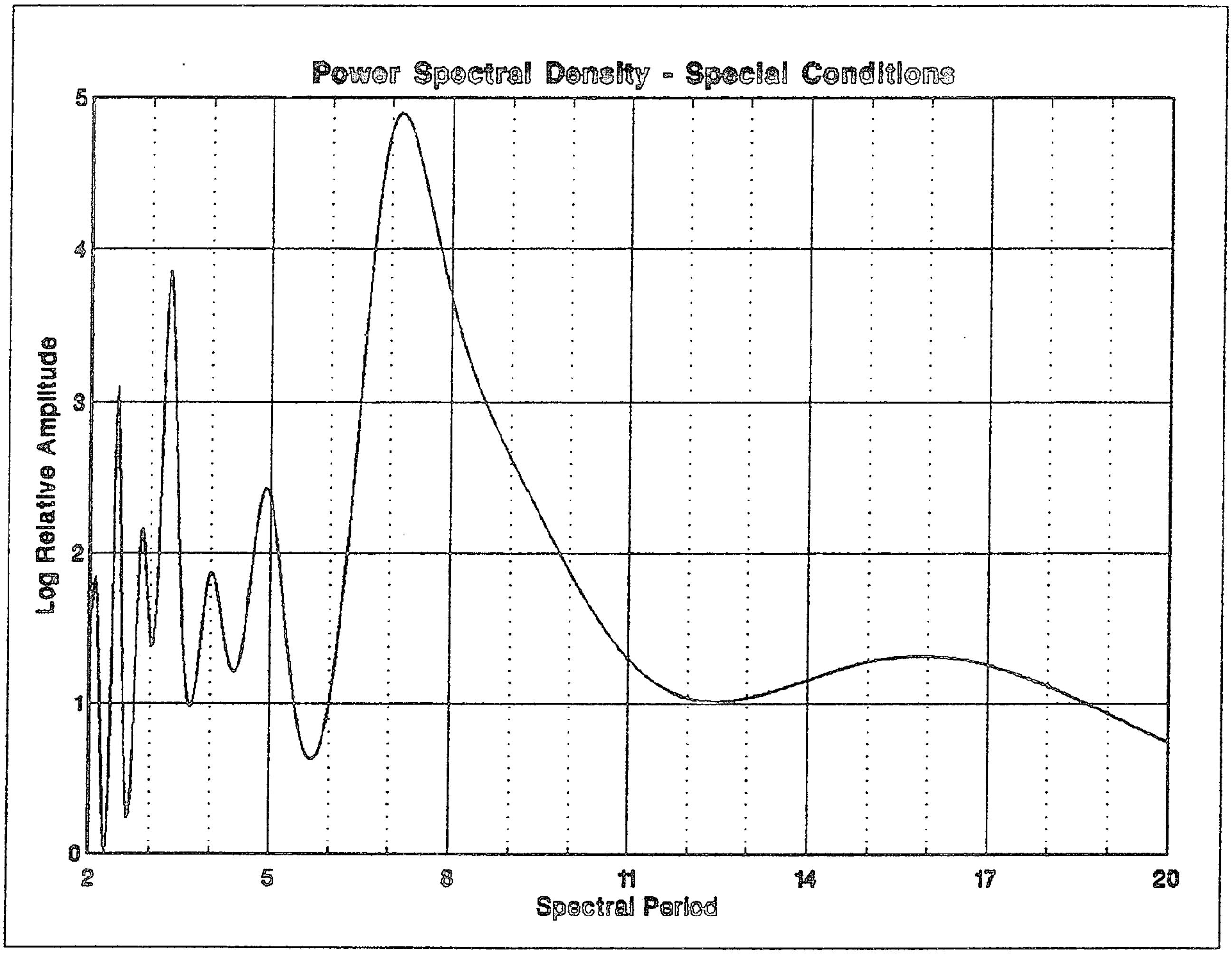

FIGURE 52. Spectrum for special conditions.

4.6 Sample Size Allocation

Figure 53 gives the makeup of the multivariate sampling strata. All ancillary variables listed within each stratum are indistinguishable from one another as predictors of DNBI incidence rates. Figure 54 is the same except for univariate sampling strata. The reason multivariate and unvariate strata are different is that the sources of variability are different. The multivariate strata are based strictly on ICD9-CM and month classes whereas the univariate strata are based on individual ICD9-CM categories and month classes.

Figure 55 shows the currently estimated sample sizes for multivariate sampling. The sample sizes represent the number of patient encounter records that must be collected each month in order to detect differences of 2.0 among DNBI incidence rates between units and between times.

MULTIVARIATE STRATUM DEFINITIONS

BRANCH Stratum 1	BRANCH Stratum 2
USN	USMC

PLATFORM Stratum 1	PLATFORM Stratum 2
LST	Ashore
ARS	AE
LPH	AFS
AOE	AR
LHA	AS
FF	AOR
CVN	AVT
LPD	CG
CGN	DDG
CV	DD
LSD	LCC
	AGF
	FFG
	LKA
	MSO

LOCATION Stratum 1	LOCATION Stratum 2
Med	Per-Red
Atl-ne	
Pac-w	
Pac-cen	
Atl-nw	
Pac-ne	

PAYGRADE Stratum 1	PAYGRADE Stratum 2
Enlisted	Officer

FIGURE 53. Multivariate stratum definitions. Inhabitants of each stratum are listed separately for each ancillary variable. Variables with insufficient information are not included in the listings.

MULTIVARIATE STRATIFICATION ANALYSIS (Number per Month)

Totals:

Variable	Total	Strata...	
BRANCH	6144	4869	1275
PLATFORM	6063	3152	2911
LOCATION	3848	3847	1
PAYGRADE	6124	6048	76

Standard Deviations:

Variable	Total	Strata...	
BRANCH	86.64	108.33	56.31
PLATFORM	74.83	137.53	46.03
LOCATION	132.23	132.27	.11
PAYGRADE	147.76	149.62	5.47

Sample Size Proportions:

Variable	Strata...	
BRANCH	.88016786	.11983214
PLATFORM	.76384417	.23615583
LOCATION	.99999974	.00000026
PAYGRADE	.99954050	.00045950

Sample Size Allocations:

Variable	Strata...	
BRANCH	3773	514
PLATFORM	3275	1012
LOCATION	4287	0
PAYGRADE	4285	2
Total	4287	

FIGURE 54. Multivariate sample size stratifications. Backup information relating to variability and stratum size are included.

UNIVARIATE STRATUM DEFINITIONS

BRANCH Stratum 1	BRANCH Stratum 2
USN	USMC

PLATFORM Stratum 1	PLATFORM Stratum 2
LSD	AOE
AS	AOR
LCC	LPD
ARS	AVT
DD	LPH
MSO	FF
CVN	AGF
CV	LST
FFG	DDG
CG	Ashore
CGN	AFS
LHA	AE
	LKA
	AO

LOCATION Stratum 1	LOCATION Stratum 2
Med	Indian
Pac-w	Atl-nw
Pac-cen	
Car-GoM	
Atl-ne	
Pac-ne	

PAYGRADE Stratum 1	PAYGRADE Stratum 2
Enlisted	Officer

FIGURE 55. Univariate stratum definitions. Inhabitants of each stratum are listed separately for each ancillary variable. Variables with insufficient information are not included in the listings.

4.7 Sampling Plan Validation

The ability of the sampling plan to detect differences of 2.0 between means was tested on artificial populations with selected ratios between mean DNBI incidence rates and with the variability of the historical data. Two separate populations were constructed and used as the basis for testing. One population was set up for sampling service branch strata and the other for sampling platform strata. Each population was used in independent tests of sampling plan validity.

For the validity tests, each population was divided into two subpopulations. Based on a total population size of 1,000,000, each subpopulation was initialized with a specified number of incidences. The number of incidences were specified to create a known (or true) ratio of incidence rates between the subpopulations. For instance, for a ratio of of differences between means equal to 1.0, 500,000 incidences were placed in each subpopulation. Each subpopulation represented a sampling stratum. Therefore, service branch and platform were divided into two subpopulations since each contains two sampling strata.

Within subpopulation variability was derived from the distribution of incidences among months of the year. The distribution of incidences was calculated independently for each sampling stratum (i.e., each subpopulation) and was based on the historical data. Specifically, for each subpopulation, the relative frequency of incidences was calculated as shown in Table 5.

TABLE 5. Simulated Population Sample Proportions

	Branch		Platform	
Subset	Subpop 1	Subpop 2	Subpop 1	Subpop 2
1	0.1017	0.0935	0.1230	0.0745
2	0.0944	0.0695	0.1121	0.0636
3	0.1126	0.0942	0.1423	0.0726
4	0.1066	0.1898	0.1354	0.1113
5	0.1039	0.1189	0.1157	0.0978
6	0.1005	0.1200	0.1080	0.1021
7	0.0737	0.0446	0.0600	0.0765
8	0.0714	0.0434	0.0569	0.0757
9	0.0585	0.0465	0.0358	0.0773
10	0.0672	0.0725	0.0451	0.0938
11	0.0630	0.0730	0.0392	0.0933
12	0.0466	0.0342	0.0265	0.0615

The result was to create 12 potential subsets of each subpopulation from which to draw samples. Therefore, each population of 1,000,000 was apportioned among 24 subsets according to the known ratio between subpopulations 1 and 2 and the relative frequencies within each. For example, given a known ratio of 1.0 with 500,000 incidences in each subpopulation, subset 1 would contain 50,850 incidences in subpopulation 1 and 46,750 incidences in subpopulation 2. For purposes of the validation, subpopulation 1 is never larger than subpopulation 2. It can be shown that continued sampling, where the subset is chosen randomly on each trial, will reproduce the variability of the original historical data.

The first step in testing was to choose a true ratio for the population. From this ratio, the test population was set up as outlined above.

The second step was to take 4287 samples (see Figure 55) from the population, each sample resulting in the recording of one DNBI incidence. For each sample, a subpopulation subset was randomly chosen and then a single incidence recorded for either subpopulation 1 or subpopulation 2. The choice of subpopulation 1 or 2 was made randomly based on the relative number of incidences in each subpopulation. For example, if subset 1 were chosen, there would be a slightly higher probability of recording the incidence in subpopulation 1 than in subpopulation 2. After the incidence was recorded, the appropriate subpopulation was decremented by one to reflect sampling without replacement. Sampling without replacement is required in order to simulate removal of a DNBI patient from the susceptible population and therefore from further epidemiological consideration. To complete this step, the numbers of incidences recorded in each subpopulation were compared to determine if $X_1 \leq 2X_2$.

The final step in testing was to repeat the second step 1000 times, counting the number of times that $X_1 \leq 2X_2$. The object was to test for the ability of the sampling plan to detect differences greater than two only when they actually occurred.

Table 6 gives the results of the validation testing. Ratios ranging from 1.0 to 3.0 were tested; values where important changes occurred are presented. The results indicate that the sampling plan is quite adequate in detecting the required differences. The error rate does not exceed 10% within an interval of $\pm$ 0.05 of the true ratio.

TABLE 6. Results of Tests on Simulated Population

Ratio	Number of Occurrences: Branch $X_1 \leq 2X_2$	Branch $X_1 > 2X_2$	Platform $X_1 \leq 2X_2$	Platform $X_1 > 2X_2$
1.00	1000	0	1000	0
1.50	997	3	999	1
1.90	951	49	918	82
1.95	942	58	902	98
1.96	850	150	634	366
1.97	618	382	570	430
1.98	577	423	520	480
1.99	554	446	502	498
2.00	510	490	499	501
2.01	507	493	501	499
2.02	473	527	489	511
2.03	299	701	383	617
2.04	155	845	181	819
2.05	40	960	90	910
2.10	9	991	30	970
2.50	11	989	2	998
3.00	0	1000	1	999

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	JULY 1992	FINAL / JAN 1992 - JAN 1993

4. TITLE AND SUBTITLE
DEVELOPMENT OF A SAMPLING STRATEGY FOR DISEASE AND NON-BATTLE INJURY (DNBI) DATA RATES

5. FUNDING NUMBERS
Program Element: 63706N
Work Unit Number:
M0095.005-6103

6. AUTHOR(S)
IVAN SHOW, Ph.D. AND MARTIN R. WHITE, M.P.H.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

8. PERFORMING ORGANIZATION
Technical Document 95-2B

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 2
Bethesda, MD 20889-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report is the user documentation for the integrated system of software programs called Epidemiological Interactive System (EPISYS). The program consists of a number of use-callable modules, EPILIMIT, EPIBASE, EPISAM, EPIMIPS, and Utilities module. The system when complete, will give researchers and medical planners the capability to easily detect, sample, and analyze any ICD9-CM illness based upon a number of ancillary variables which include age, race, sex, service branch, ship type, pay grade, and occupation. These programs will significantly improve the ability to access our Medical History Files, and will provide investigators with an integrated system of computer programs for health monitoring and medical projection needs. The user documentation in this report explains in detail the user module, EPISAM.

14. SUBJECT TERMS
EPIDEMIOLOGY
MEDICAL SURVEILLANCE
COMPUTERIZED MEDICAL INFORMATION SYSTEMS

15. NUMBER OF PAGES
90

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited